Agrupamento de contextos de palavras polisémicas

Luís Sarmento
las@fe.up.pt

Resumo

Neste relatório iremos apresentar algumas experiências de aplicação de técnicas de agrupamento com o objectivo de determinar os possíveis sentidos de uma palavra polisémica. Serão comparados 5 algortimos de agrupamento no processamento de informação acerca das co-ocorrências entre palavras obtidas de uma colecção de 1000 milhões de palavras. Serão também analisados os impactos da alteração de vários parâmetros relativos às medidas de associação empregues e relativos à escolha dos sub-conjuntos de dados considerados.

1 Introdução

A existência de palavras polisémicas introduz na linguagem um factor de ambiguídade que dificulta a sua análise automática. Por exemplo, na pesquisa de elementos semelhantes à palavra "laranja" ficamos imediatamente confrontados com a possibilidade distinta de escolher - pelo menos - entre elementos do conjunto das cores ou dos frutos, não sendo possível sem informação contextual adicional decidir quais os elementos "verdadeiramente" semelhantes que se pretende encontrar. O problema de decidir entre quais dos sentidos de uma determinada palavra polisémica está a ser referênciado num determinado contexto é conhecido como "desambiguação de sentidos" e representa actualmente uma área de investigação muito activa.

Como notado em (Yarowsky, 1993) as palavras polisémicas apresentam grupos de colocações distintas em função dos seus diferentes sentidos, o que abre a possibilidade para efectuar a desambiguação usando a informação relativa a essas mesmas colocações. Por exemplo, tomando o exemplo da palavra polisémica "planta", que pode ter pelo menos os sentidos (i) de vegetal, (ii) de elemento cartográfico/arquitectónico ou (iii) como forma abreviada do termo anatómico "planta dos pés", poderiamos eventualmente encontrar em corpora colocações como "podar a planta", "ler a planta" ou "massajar a planta" que nos permitem distrinçar entre os diferentes sentidos associados.

Neste trabalho pretendemos explorar algumas questões associadas à desambiguação de sentidos e à descoberta de possíveis contextos, tentando verificar até que ponto se torna possível aproveitar eficientemente a informação acerca das suas co-ocorrências. Procuraremos também relaxar a condição relativa à existência de colocações típicas, que normalmente requerem um certo pré-processamento linguístico na sua identificação, para considerar somente a informação acerca das palavras com as quais uma determinada palavra polisémica co-ocorre no âmbito de uma unidade de discurso bem delimitada (no nosso caso uma frase). A ideia base que pretendemos explorar é a de que, para um determinado sentido, uma palavra polisémica deverá co-ocorrer com um conjunto de palavras típico desse sentido, e estas, por reflexão deste racicionio, deverão também co-ocorrer entre sí nesse mesmo determinado contexto, gerando eventualmente agrupamentos coesos de palavras. Tendo em conta que as palavras co-ocorrem normalmente com muitas centenas ou milhares de outras palavras, esta parece ser uma boa oportunidade para a experimentação de técnicas de "data-mining", em particular de agrupamento, já que o que pretendemos é descobrir os agrupamentos típicos, usando vastas quantidades de informação. Iremos fazer uso do pacote de "data-mining" R, distribuido sob licença GPL em http://www.r-project.org/, que contém vários algoritmos de agrupamento que poderão assim ser usados de uma forma mais simples. A informação acerca das co-ocorrências que pretendemos explorar será obtida a partir do BACO, uma base de dados de co-ocorrências

compilada a partir da colecção de documentos Web WPT03 (Gomes et al., 2006), sendo esta também uma oportunidade para tentar aferir o potêncial deste recurso.

2 Trabalho Relacionado

O nosso trabalho está intimamente relacionado com a desambiguação de sentidos de uma palavra polisémica. Sobre esta temática têm sido desenvolvidos bastantes trabalhos usando aproximações automáticas, baseadas em técnicas de aprendizagem e em particular de agrupamento. Iremos contúdo apenas referir um trabalho (Yarowsky, 1995) muito interessante que ilustra bem o interesse na exploração da informação acerca das colocações. Yarowsky apresenta um algoritmo de aprendizagem não supervisionada com o objectivo de desambiguar palavras polisémicas. O algoritmo parte do princípio que as colocações de uma dada palavra (polisémica) não são partilhadas pelos seus diversos sentidos e utiliza a informação acerca das colocações típicas para separar os respectivos sentidos possíveis. O algoritmo começa por tentar, a partir de um conjunto de frases exemplo fornecido pelo utilizador e relativas a dois sentidos possíveis da palavra (no artigo são assumidos apenas dois sentidos possíveis por palavra), inferir um conjunto de regras acerca das colocações que permitem separar referidos sentidos. Usando essas regras, inicia-se então um agrupamento das restantes frases concordantes com a palavra polisémica, tentando-se obter, por aglomeração de novos casos (frases), mais colocações típicas de cada sentido. Com as novas colocações vai sendo possível refinar as regras iniciais e obter regras adicionais de classificação baseadas unicamente no conjunto e na distância das colocações observadas. Estas regras permitem assim determinar o sentido da palavra polisémica em causa em frases não observadas ou não desambiguadas anteriormente, sendo que o algoritmo termina quando o processo converge (pode permanecer um resíduo de frases não agrupadas/desambiguadas). Apesar de ser um algoritmo de aprendizagem de regras de classificação, este algoritmo possui muitas das caracteristica base dos algoritmos de clustering aglomerativos, já que tenta algomerar para cada sentido da palavra polisémica um conjunto de novos exemplos e colocações usando uma métrica de distância relacionada com a verificação de uma regra colocacional.

Nos pretendemos seguir a pista deste trabalho explorando informação acerca das co-ocorrências entre palavras, e usando algoritmos de agrupamento standard disponibilizados publicamente no pacote R.

3 Os dados

A informação base que usamos para as nossas experiências de clustering encontra-se aramzenada no BACO. O BACO é uma base de dados gerada a partir de colecção de textos WPT03 e que armazena não só o texto proveniente dessa colecção mas também informação relativa a sequências de 1, 2, 3 e 4 palavras consecutivas e das co-ocorrências entre pares de palavras. Após remoção de duplicados, a colecção WPT03 contém cerca de 1000 milhões de palavras pelo que se pode admitir que o BACO contém uma quantidade de informação suficientemente grande para permita a aplicação de métodos intensivos de análise de dados.

A informação do BACO com a qual iremos lidar directamente é a tabela de co-ocorrências, que armazena tuplos da forma (p1,p2,f). Estes tuplos armazenam a informação de que a palavra p1 ocorre antes da palavra p2 com frequência f, no universo das frases do WPT03. A tabela das co-ocorrências armazena 780M de tuplos, correspondendo assim a um grafo dirigido com 780 milhões de arestas e cerca de 4 milhões de nós.

Esta tabela permite obter com relativa rapidez, normalmente entre 1 a 15 segundos a informação relativa a todas as co-ocorrências de uma determinada palavra, possibilitando assim obter os dados relevantes para as experiências de agrupamento que pretendemos efectuar. Para uma determinada palavra polisémica, poderemos então obter a lista de palavras com as quais esta palavra co-ocorre. Numa segunda fase poderemos obter a frequência de co-ocorrências entre estas. Produz-se assim uma matriz, $FM(p_i)$, que armazena o valor da frequência de co-ocorrência mútuas, ou seja, entre as palavras que co-ocorrem com a palavra p_i.

Seja $C(p_i)$ o vector de co-ocorrências da palavra p_i , e seja e $FC(p_i)$ o vector contendo o valor da frequência das co-ocorrências da palavra p_i:

$$C(p_i) = \{p_j, p_k, p_l, p_m, p_n \ldots\}$$

$$FC(p_i) = \{fc_i(p_j), fc_i(p_k), fc_i(p_l), fc_i(p_m), fc_i(p_n) \ldots\}$$

O termo $fc_i(p_j)$ refere-se ao valor da frequência da co-ocorrência ente as palavras p_i e p_j, obtido por consulta ao BACO. Para cada palavra co-ocorrente com pi, isto é p_k _ $C(p_i)$ torna-se possível obter o vector FCM_k com os valor da frequência das co-ocorrências mútuas de p_k, ou seja entre p_k e todas os outros elementos de $C(p_i)$:

$$FCM_k = \{fc_k(p_j), IND, fc_k(p_l), fc_k(p_m), fc_k(p_n) \ldots\}$$

O valor de $fc_k(p_k)$, ou seja da co-ocorrência de p_k consigo mesmo, é considerado indefinido. Os vários vectores FCM_k podem ser organizados numa matriz simétrica $MFM(p_i)$ da forma:

$$MFM(p_i) = \begin{matrix} FCM_j \\ FCM_k \\ FCM_l \\ FCM_m \\ \ldots\ \ldots\ \ldots \end{matrix}$$

Esta será a matriz que contém a informação base acerca das co-ocorrências mútuas das palavras que co-ocorrem com a palavra inicial p_i (como referido anteriormente o valor de $fc_k(p_k)$ é considerado indefinido para evitar a tendência de forte diagonalização da matriz).

4 Distância e Medidas de Associação

Os valores da frequência que se retiram do BACO devem ser considerados informação em bruto: referem-se apenas à frequência de co-ocorrência entre duas palavras e não servem só por sí para estabelecer medidas fieis proximidade / coesão entre palavras. Por exemplo, certos adjectivos, ou verbos frequentes apresentam frequências de co-ocorrências muito elevadas com grande parte do léxico sem que isso signifique que possuam um grau de coesão com um palavra em particular com a qual co-ocorram com uma razoável frequência. Adicionalmente, duas palavras que co-ocorram com grande frequência com o verbo "ser" não podem ser consideradas, de alguma forma, semelhantes só por isso. Por consequência, os valores da frequência obtidos também não podem servir de base para critérios de semelhança vectorial necessários para os algoritmos de agrupamento que iremos utilizar.

A noção de proximidade entre palavras, está intimamente relacionada com a noção de medida de associação, área na qual têm sido desenvolvidos vários trabalhos - para uma visão geral ver (Evert, 2005) e o (Schone and Jurafsky, 2001). Para o nosso trabalho iremos utilizar a medida de associação Informação Mútua (Church and Hanks, 1990), que é relativamente tradicional no processamento estatístico da linguagem, embora seja aplicável genericamente a vários domínios.

A Informação Mútua (IM) é uma medida que mede o grau de associação entre dois acontecimentos e têm como base a relação que existe entre a probabilidade da co-ocorrência dos dois acontecimentos relativamente à probabilidade da sua ocorrência individual. A IM entre dois acontecimentos x e y é dada por:

$$I(x;y) = \log_2 P(x,y) / (P(x) P (y))$$

Se os dois acontecimentos forem estatisticamente independentes então:

$$P(x,y) = P(x) P (y)$$

e logo o valor de I(x;y) será zero. Por outro lado, quanto maior for a dependência entre os dois acontecimentos maior será P(x,y) relativamente ao produto das probabilidades individuais e por isso o valor da função I(x;y) será também maior.

Transportando a IM directamente para o nosso caso, P(x,y) representará a probabilidade da co-ocorrência das palavras x e y na colecção WPT, enquanto que P(x) e P(y) representará simplesmente a probabilidade da ocorrência dessas palavras em qualquer situação. Os dados que podemos extrair do BACO são relativos às frequências e não às probabilidades em sí. Por isso, teremos de estimar o valor da função de probabilidade, tendo em conta os dados da frequência e do número de palavras armazenadas no BACO. Sabemos que o a função de probabilidade pode ser estimada a partir do valor das frequências da seguinte forma:

$$P (x) \sim F(x) / N$$

sendo N o somatório das frequências de todas as ocorrências. No nosso caso N será o número de palavras do WPT armazenadas no BACO, pelo que o seu valor é exactamente de 1.059.436.086. A Informação Mútua pode então ser estimada por em função dos valores das frequências:

$$I(x;y) \sim \log_2 \frac{F(x,y)/N}{(F(x)/N)\,(F(y)/N)} = \log_2 \frac{N\ F(x,y)}{F(x)\ F(y)}$$

Será esta a fórmula usada para obter os valores que servirão de base ao agrupamento. Assim, os algoritmos de agrupamento aos quais recorreremos, não irão operar sobre a matriz de frequências $FM(p_i)$ apresentada na secção anterior, mas sim sobre uma matriz análoga em que os valores das frequências de co-ocorrência entre duas palavras são substituidos pelo valor da informação mútua calculada pela fórmula anterior. Essa será a matriz $MIM(p_i)$, Matriz de Informação Mútua, composta pelos valores da Informação Mútua entre os elementos que co-ocorrem com a palavra p_i.

5 A Informação Mútua e a Ponderação adicional

É sabido que formula da Informação Mútua tem tendência a promover acontecimentos raros, ou seja, assume valores superiores quando P(x) e ou P(y) possuem valores muito reduzidos. Por isso, e para evitar a promoção tendenciosa de certas ocorrências mais raras, tem vindo a ser sugeridas algumas formúlas de compensaçãodeste efeito adverso. Para o nosso trabalho não iremos contudo usar nenhuma destas fórmulas mas tentaremos reduzir os efeitos tendenciosos excluindo da matriz MIM palavras muito pouco frequentes e que por isso levariam a valores de Informação Mútua indesejavelmente altos.

No entanto, para além de usarmos a Informação Mútua, que é uma medida de associação standard, incluímos nas nossas experiências algumas variantes empíricas que foram usadas de forma a poder testar a importância relativa de vários factores que consideramos relevantes.

Em primeiro lugar decidimos introduzir a posibilidade de se usar a medida de associação mais elementar, isto o valor da frequência da co-ocorrência. Ou seja, mesmo sabendo as fragilidades de uma medida de associação baseada unicamente neste valor julgamos que seria importante ainda assim poder aferir na prática os problemas que dela advêm.

Em segundo lugar introduzimos também um factor de ponderação opcional e aplicável às duas medidas de associação anteriores, e que tenta integrar no processo de agrupamento a informação relativa ao grau de associação entre a palavra inicial e as suas co-ocorrentes. Recorde-se que a matriz relativa às co-ocorrências mútuas, $MIM(p_i)$, considera apenas a informação de co-ocorrência entra os pares de palavras co-ocorrentes e não recolhe nenhum factor que traduza a influência que possa existir entre qualquer uma dessas palavras co-ocorrentes e a palavra inicial. Por isso, introduzimos um factor de ponderação que multiplica cada elemento da matriz $MIM(p_i)$, pelo valor da Informação Mútua entre os respectivos elementos co-ocorrentes e a palavra p_i. Consegue-se assim introduzir um factor extra (opcional) que permite incluir a influência do grau de associação de cada uma das palavras co-ocorrentes com a palavra inicial.

6 Aplicação desenvolvida

A aplicação desenvolvida terá, por uma lado, de recolher os dados armazenados no BACO e, por outro, terá de invocar os algortimos de agrupamento acessiveis a partir do pacote aplicacional R. Por uma questão de comodidade, a aplicação que construimos foi desenvolvida em Perl. O acesso ao BACO pode ser feito de uma forma muito simples usando o módulo Perl BACO.pm, enquanto que para fazer o interface com a o pacote R existe a possibilidade de instalar um módulo de interface específico. Contúdo, e após várias tentativas, não conseguimos fazer a instalação do referido módulo de interface com R, por uma questão de incompatibilidade com a versão Perl instalada. Por esse motivo, a comunicação entre o nosso programa Perl e o pacote R foi realizada por intermédio de ficheiros e chamadas a linhas de comandos construídas dinamicamente.

Resumidamente, o fluxo do programa desenvolvido é o seguinte:

1. Para uma dada palavra fornecida como parâmetero, consultar o BACO para obter a lista de palavras co-ocorrentes.
2. Para cada uma das palavras co-ocorrente obter a sua frequência e para aquelas que respeitem critérios mínimos de frequência (fornecidos como parâmetro), obter as lista completa das suas co-ocorrências, que assim ficam indexadas na memória principal.
3. Para cada uma das palavras co-ocorrentes selecionadas, consultar a lista completa das suas co-ocorrências para determinar os valores de frequência de co-ocorrência com as restantes. Desta forma consegue-se obter um vector de frequências mútuas que é transformado num vector com os valores de Informação Mútua usando os parâmetros de frequência anteriormente obtidos. Depois de processar todos os elementos co-ocorrentes obtem-se a matriz que servirá de base para o agrupamento.

4. Com os valores calculados, e com os parâmetros relativos ao agrupamento fornecidos pelo utilizador, construir um script R e guardá-lo num ficheiro.
5. Chamar uma linha de comando do sistema para executar o R com o script gerado anteriormente. Guardar os resultados produzidos em ficheiro.
6. Ler os resultados do ficheiro e apresentar ao utilizador.

O fluxo é simples sendo que a maior parte do tempo de computação encontra-se na fase 2 (acesso intensivo ao BACO) e 5 (execução do algoritmo de agrupamento via pacote R).
A aplicação recebe vários parâmetros que permitem por um lado controlar a informação que é considerada de toda aquela que é extraída do BACO, e por outro, manipular toda uma série de variáveis associadas a processo de clustering. Os parâmetros recebidos são:

- **p**: palavra p_i (possivelmente) polisémica em torno da qual se fará a análise
- **max** (inteiro): número máximo de palavras co-ocorrentes a considerar.
- **minfreq** (inteiro): frequência mínima para que uma palavra co-ocorrente seja incluida na matriz MIM(p_i).
- **mincooc** (inteiro): frequência de co-ocorrencia mínima com p_i para que uma dada palavra seja incluida na matriz MIM(p_i).
- **medida**: métrica utilizada para definir a associação entre palavras no cálculo da matriz MIM(p_i). Opções possíveis: Informação Mútua ou Frequência de Co-ocorrências. Opção adicional de se introduzir factor extra de ponderação.
- **alg**: qual o algoritmos de clustering a ser invocado via R. Opções: DIANA, AGNES, CLARA, PAM e FANNY.
- **norm** (booleano): opção de normalização adicional da matriz MIM(p_i), a ser executada via R.
- **nclusters** (inteiro): número de clusters solicitado ao algoritmo de agrupamento selecionado.
- **baco**: número IP da base BACO a consultar.
- **verbose** (inteiro): nível de produção de mensagens de aviso durante o processo.

7 Algoritmos de agrupamento utilizados

Para este trabalho utilizamos unicamente os algoritmos de agrupamento disponíveis no Pacote R. Uma parte significativa do esforço de desenvolvimento esteve relacionada com a compreensão dos parâmteros associados a cada algoritmo, à forma como podem ser invocados e à forma como os seus resultados podem ser lidos. Em seguida descrevemos sucintamente os algoritmos disponíveis no pacote R e com os quais realizamos alguma experimentação. A maior parte da informação que conseguimos recolher provém da documentação do próprio pacote R e de (Berkhin, 2002).

7.1 DIANA

O algortimos DIANA (DIvisive ANAlisys) é um algoritmo de agrupamento hierárquico divisivo que começa por considerar um único agrupamento contendo todos os casos, e executa divisões hierárquicas sucessivas até obter agrupamentos com um elemento apenas. O resultado do algortimo DIANA é assim um dendograma de altura mínima $\log_2$(número de casos), para o caso de um dendograma totalmente equilibrado - situação que nas nossas experiências nunca aconteceu, bem pelo contrário.
Em cada iteração, é escolhido o agrupamento de maior dimensão para ser divido em dois agrupamentos mais pequenos. O diâmetro de um agrupamento é simplesmente o valor da maior distância entre qualquer par dos seus elementos (o que obriga à comparação de todos os pares). A divisão é realizada em torno do elemento que possui um maior afastamento relativamente a todos os outros elementos do agrupamento inicial, que assim inicia um novo grupo. Este novo agrupamento passa a absorver todos os elementos do grupo inicial que se encontrem mais próximos dele que do grupo inicial, permitindo assim a criação de dois novos agrupamentos.

7.2 AGNES

O algoritmo AGNES é também um algoritmo hierárquico embora com uma estratégia aglomerativa (ou seja, inversa ao do DIANA). No algoritmo AGNES, o estado inicial inclui tantos grupos quantos os elementos a agrupar. Iterativamente, os agrupamentos vão sendo fundidos gerando agrupamentos cada vez maiores até todos os elementos fazerem parte de um único agrupamento. Em cada iteração, são

fundidos os dois agrupamentos mais próximos sendo que a distância entre dois agrupamentos pode ser calculada de várias formas:

- média: a distância entre dois agrupamentos é a média das distâncias entre os pontos de cada um deles.
- singular: a distância entre dois agrupamentos é a menor distância estabelecida entre dois pontos pertencentes a cada um do agrupamentos.
- completa: a distância entre dois agrupamentos é a maior distância estabelecida entre dois pontos pertencentes a cada um dos agrupamentos.

Existem ainda fórmulas adicionais para o cálculo de distâncias embora sejam mais complexas e por isso não estão no âmbito deste pequeno resumo. O resultado do algoritmo AGNES é também um dendograma, do qual se podem retirar agruapmentos cortando a árvore em diferentes níveis.

7.3 PAM

O algoritmos PAM (Partitioning Around Medoids) tenta encontrar objectos representativos - os medoides - entre o conjunto de elementos a agrupar, em torno dos quais se deverão agrupar os restantes elementos. Depois de conseguir encontrar k medoides, que formarão um conjunto de agrupamentos inicial, o algoritmo tenta simplesmente atribuir os restantes elementos ao agrupamento mais próximo. A ponto crucial do algoritmo consiste em encontrar os medoides que são mais representativos de entre todos os elementos a agrupar. O algoritmo começa por tentar procurar um conjunto inicial de medoides "bons". Em seguida inicia um processo de optimização efectuando várias trocas entre medoides actuais e outros elementos no sentido de minimzar uma determinada função objectivo. Quando se encontra um mínimo (local?), os medoides actuais são considerados definitivos e faz-se o agrupamento tendo em conta as distâncias relativas a esses medoides.
Este algoritmo tem no entanto como principal inconveniente a dificuldade em escalar para grandes conjuntos de dados já que o problema de optimização associado é algoritmicamente complexo. No manual associado ao pacote R é apontado como referência o número de 200 elementos, como um limite máximo adequado para este algoritmo. A nossa experiência concreta diz-nos que este valor pode ser duplicado sem que isso represente um problema de performance, mesmo para um espaço de atributos com dimensão na ordem das centenas.

7.4 CLARA

O algoritmo CLARA (Clustering LARge Applications) tenta resolver o problema de escalabilidade do algoritmo anterior realizando uma amostragem prévia de dados antes de iniciar a pesquisa dos medoides. De entre o conjunto inicial a agrupar, é retirada uma amostra, que se espera representativa, e sobre essa é feita a pesquisa de medoides de uma forma semelhante ao PAM. Com isto espera-se reduzir o problema de escalabilidade associado ao PAM. No sentido de reduzir o risco de uma escolha muito negativa da amostra inicial, o CLARA repete o processo completo várias vezes e retornado apenas o melhor agrupamento resultante, verificando a sua qualidade através da medição das distâncias entre os medoides relativamente a todos os elementos a agrupar (e não apenas os amostrados). A desvantagem deste algortimo prende-se com o facto de que um agrupamento realizado com a amostragem inicial não garante por sí um bom agrupamento final com a totalidade dos elementos a agrupar.

7.5 FANNY

O algoritmo FANNY (Fuzzy Analisys Clustering) é um algoritmo distinto dos anteriores já que prevê a possibilidade de um determinado elemento poder fazer parte de vários agrupamentos com um determinado grau de pertença. Os graus de pertença de um determinado elemento aos vários conjuntos possíveis variam entre 0 e 1, sendo o somatório dos graus de pertença de cada elemento igual a 1.
Não foi infelizmente possivel obter mais informação relativamente ao modo de funcionamento deste algoritmo. Mas pela sua experimentção foi possível verificar que este algoritmos (pelo menos a implementação acessivel via o pacote R) é diferente de todos os anteriores por ser capaz de se adaptar ao dados e produzir um número de agruapmentos diferentes do solicitado pelo utilizador como parâmetro inicial. Todos os outros algoritmos, produzem um número de grupos exactamente igual ao solicitado pelo utilizador, o que por vezes é problemático de definir sem alguma experimentação, enquanto que FANNY, em todas as nossas experiências demonstrou ser capaz de produzir automaticamente um número

de agrupamentos inferior ao solicitado. Desta forma, o FANNY permite ao utilizador definir um majorante para o número de agrupamentos desejado, sendo o número de agrupamentos do resultado final um valor ajustado automaticamente em função dos dados fornecidos.

8 Experiências Efectuadas

O programa desenvolvido, possui muitos parâmetros, pelo que a sua exploração é um processo exaustivo e demorado. Pensamos que é nesta fase mais importante poder perceber o impacto de alguns destes parâmetros nos resultados finais do que procurar encontrar uma combinação óptima que leve ao melhor agruapamento possível. De facto, as variáveis em jogo são tantas que impõem uma análise o mais ortogonal possível, mesmo sabendo que isso teoricamente pode ser impossível. Tentaremos assim experimentar separadamente alguns dos parâmetros envolvidos. Toda a avaliação será feita informalmente, e avaliaremos o impacto dos parâmetros tentando apenas verificar o sucesso construção de agrupamentos inteligíveis.

8.1 Algoritmos de Agrupamento

Começaremos por fixar todos os parâmetros à excepção da escolha do algoritmos de agrupamento. Iremos tentar verificar quais os agrupamentos que se criam em torno da palavra "planta". Tentaremos obter 400 palavras co-ocorrentes, considerando apenas válidas aquelas que co-ocorrem pelo menos 50 vezes e que ocorram na colecção pelo menos 250 vezes, de forma a evitar a promoção de casos demasiado raros. Usaremos a medida de associação da informação Mútua e não iremos introduzir o factor de ponderação que descrevemos na secção 5. Iremos também solicitar ao algoritmo a construção de um número razoavelmente grande de grupos, com o objectivo de verificar se são criados grupos verdadeiramente coesos (com város elementos) enquanto que os elementos menos coesos ficarão isolados num grupo próprio. Assim, pretendemos criar grupos com várias palavras (mais do que 5-10), apesar de consideramos a possibilidade de vários elementos ficarem isolados, por exemplo 1 ou 2 por cada 10 palavras. Como heurística iremos permitir a criação de grupos num número de de 15 a 20% do máximo de palavras co-ocorrentes permitido (400). Decidimos fixar o valor em 75 grupos. Este pode parecer um valor excessivo, mas pelo facto da nossa aproximação tentar o agrupamento contextos, que são em grande número para cada sentido da palavra, este valor revela-se apropriado, especialmente tendo em conta a presença de palavras ruídosas que gostariamos de conseguir isolar.

Os resultados apresentados em seguida reportam-se à mesma matriz de co-ocorrências sendo que a única condição que se altera é a referente ao algoritmo de agrupamento empregue.

Resultados com DIANA (Anexo A.1)

Os resultados obtidos com este algoritmo foram animadores. Foi possível obter 30 agrupamentos com 5 ou mais elementos, sendo que na maior parte destes é possível identificar com razoável facilidade que se tratam ou contextos associados ao sentido de planta como "vegetal" ou "planta arquitectónica". É também possível verificar a existência de alguns agrupamentos com 2, 3 ou 4 palavras que também pertecem claramente a contextos bem definidos de um dos sentidos da palavra "planta", como por exemplo ("topográficos", "levantamentos"). Outra coisa que ressalta deste agrupamento é que a grande maioria das palavras consideradas aparentam estar relacionadas com o sentido arquitectónico da palavra planta, o que sugere que a colecção tem uma forte preponderância deste domínio ou a medida de associação usada para filtar a palavras co-ocorrentes a considerar introduz uma distorção que teremos que estudar futuramente.

Resultados com AGNES (Anexo A.2)

Os resultados da aplicação do algoritmo AGNES não parecem ser muito diferentes apresentando contudo um grau de desiquilibrio maior: apenas 17 agrupamentos possuem 5 ou mais elementos. Por inspecção é dado a entender que este algoritmo juntou vários dos agrupamentos relacionados com o sentido arquitectónico da palavra planta, e que pela utilização do algoritmo DIANA haviam permanecido separados. Esta parece ser a principal diferença entre os algoritmos, que neste caso levaram à produção de um agrupamento dominante com mais de 30% dos elementos. Esta situação não parece ser um problema intrinseco ao algortimo mas parece estar relacionada com o processo de escolha dos nós do dendograma produzido. O processo de leitura do dendograma não é simples e usando os métodos fornecidos pelo pacote R somos obrigados a confiar num corte do dendograma que parece não ser o mais

apropriado. Infelizmente, não encontramos alternativa simples ao método de corte fornecido pelo R.

Resultados com PAM (Anexo A.3)

Os resultados obtidos com o algortimo PAM também não diferem muito dos anteriores. Há talvez um aparente aumento de inteligibilidade em alguns agrupamentos, sendo que em alguns casos se obtêm alguns cuja relação com a palavra planta não aprece muito evidente embora os elementos manifestem alguma coesão semântica. Por exemplo, foi encontrado os agruapamentos ("quintã", "celorico", "mogadouro", "beira"), ("lageosa", "calvário", "soeiro", "linhares") ou ("gare", "atalaia", "viçosa", "mangualde") que unem elementos que se referem nitidamente a topónimos. Parece também terem sido criados alguns agrupamentos muito coesos, este sim com uma relação mais próxima aos dois sentidos da palavra planta que temos vindo a encontrar, e que não existiam com esse nível de pureza nos casos anteirores. Destacam-se por exemplo ("pluviais", "arruamentos", "esgotos", "pavimentos", "drenagem", "abastecimento") ou ("grão", "seca", "folhas", "porte", "folha", "raiz", "raízes", "tronco", "fruto"). Um outro factor que nos parece relevante neste algoritmo é o facto de a interpretação do resultado ser feita directamente, não sendo necessário processar uma estrutura como um dendograma para obter os agrupamentos desejados.

Resultados com CLARA (Anexo A.4)

Tal como previsivel dada a relação entre os algortimos, os resultados do CLARA são bastante próximos do PAM. As diferenças visíveis nesta experiência são essencialmente a criação de um agrupamento dominante com mais de 25% das palavras. O agrupamento parece bastante coeso e relacionado com o sentido "arquitectónico" da palavra "planta", apesar possuir algumas palavras que poderiam eventualmente ser incluidas em agrupamentos relacionadas com o sentido "vegetal". É interessante verificar que esta leve tendência para agrupar mais os resultados que o PAM, levou à criação de um agrupamentos que contém mais topónimos que apenas os 4 que identificamos no exemplo anterior, nomeadamente o agrupamento: ("lageosa", "calvário", "gare", "soeiro", "linhares", "atalaia", "viçosa", mangualde"), o que no nosso entender traduz uma tendência positiva.

Resultados com FANNY (Anexo A.5)

Os resultados obtidos com o algoritmo FANNY foram, em grande medida, uma desilusão. De facto, foi apenas possível obter um grupo dominante com 327 elementos, 11 grupos com dois elementos, tendo os restantes 39 agrupamentos obtidos recolhido apenas um elemento. Destaca-se também o facto de o algoritmo não ter produzido todos os agrupamentos que foram solicitados, o que anteriormente nos pareceu uma propriedade interessante mas que, pelo menos para este caso, foi extremamente contraproducente. Como não há muita informação acerca destes algoritmos não nos atrevemos sequer a tentar esboçar uma explicação.

8.2 Medidas de Associação e factor de Ponderação adicional

Interessa também poder aferir o impacto da alteração da medida de associação utilizada para construir a matriz base para o agrupamento. Para isso, fixaremos um dos algortimos, e faremos variar exclusivamente o parâmetro relativo à medida de associação. Por motivos de eficiência computacional iremos centrar todas as nossas experiências no algoritmo de agrupamento CLARA, cuja escalabilidade pareceu-nos de facto ser a melhor, e como pudemos verificar dos testes anteriores, apresentou resultados interessantes do ponto de vista do agrupamento.

A experiência efectuada na secção anterior usou como medida de associação a Informação Mútua, e não foi incluido o factor de ponderação extra que propusemos na secção 5. Em seguida iremos gerar matrizes com as restantes 3 combinações, Informação Mútua com Ponderação, Frequência Simples e Frequência com Ponderação. Para os casos em que foi utilizado o valor da frequência é necessário proceder a uma normalização da Matriz das Frequências Mútuas, algo que é solicitado por passagem de parâmetros aos algoritmos de agrupamento.

Informação Mútua com Ponderação (Anexo B.1)

O impacto da inclusão deste parâmetro é difícil de aferir já que as diferenças entre a inclusão do factor de ponderação ou não revelam-se diminutas. Sem querer tirar nenhuma conclusão definitiva, verfica-se a "purificação" de alguns dos agrupamentos com a remoção de certas palavras ruidosas. Por exemplo, enquanto

que sem o factor de ponderação se obtiveram agrupamentos significativos mas com palavras que podem ser consideradas espúrias, como foi o caso dos agrupamentos (*quadrangular*, *quadrada*, abóbada, cúpula capela-mor, sacristia, claustro, naves) ou (caule, pigmentação, *medianamente*, caules, limbo, vegetativo, míldio, susceptibilidade, floração, polpa, parasitas) o a inclusão do factor de ponderação levou à obtenção dos agrupamentos (abóbada capela-mor sacristia claustro) e (caule pigmentação caules vegetativo floração polpa) com o afastamento de palavras como "quadrangular" e "quadrada" do primeiro agruapamento e "medianamente" do segundo. No entanto, pela comparação de todos os agrupamentos não conseguimos retirar conclusões definitivas já que, para certos casos com a inclusão do factor de ponderação verifica-se a criação de agrupamentos mais difusos que os obtidos sem o factor de ponderação. Esse é o caso, por exemplo, do agrupamento (lageosa, balcões, calvário, soeiro, moinhos, decorativos, linhares, numeração, notáveis, mondego, aldeias, janelas, religiosos, pormenores, chão, beira, mangualde, ribeira).

Frequência, sem ponderação e com normalização (Anexo B.2)

A primeira observação que ressalta imediatamente é a de que a utilização o valor da frequência como medida de associação resulta na criação de uma conjunto de agrupamentos mais desiquilibrado. Verifica-se existir uma certa tendência para a criação de agrupamentos extremos: poucos agrupamentos muito grandes (7) e muitos agrupamentos (68) com quatro ou menos elementos.

Isto não é de forma alguma só por sí um problema: se os agrupamentos tiverem um bom nível de pureza, então essa situação é até muito interessante por permitir de uma forma eventualmente mais simples verificar o número de possíveis sentidos (não contextos) de uma dada palavra. E de facto, pelo menos para esta experiência, o resultado é intrigante: apesar de terem sido construídos grupos grandes que sem dúvida confudem vários contextos dentro de um mesmo sentido, e apesar de haver nesses agrupamentos alguns elementos ruidosos, a verdade é que conseguimos facilmente dividir estes agrupamementos pelos dois sentidos que temos considerado. Mas por outro lado, não encontramos contextos tão bem delimitados como os agrupados pela medida de associação Informação Mútua, que parece criar mais grupos de dimensão intermédia, e só raramente tão grandes.

Frequência, com ponderação e com normalização (Anexo B.3)

A inclusão do factor de ponderação parece não ter alterado grandemente o panorama anterior. Aparentemente, a influência da escolha de medida de associação baseada na frequência parece dominar completamente o resultado final, e a inclusão do factor de ponderação adicional só muito marginalmente parece ter tido impacto. Contudo, parece verificar-se um efeito inverso ao encontrado com a inclusão do factor de ponderação na situação na situação em que a medida de associação utilizada era a Informação Mútua: há a criação de agrupamentos maiores e, em certos casos, bastante ruidosos. O impacto da inclusão do factor de poderação neste caso parece ter sido ligeiro mas negativo.

8.3 Impacto dos parâmetros de filtragem

As palavras consideradas durante o agrupamento resultam de um sub-conjunto de palavras co-ocorrentes com a palavra inicial que se encontram entre aquelas que apresentarem um maior valor de Informação Mútua com ela. Sabemos que a Informação Mútua tende a promover palavras raras e para evitar que fossem consideradas palavras demasiado raras foi incluindo um parâmetro que impõe uma frequência mínima às palavras para serem consideradas no agrupamento. Até agora temos obrigado a que as palavras em causa ocorram pelo menos 250 vezes na colecção, mas nesta última experiência iremos reduzir esse limite para 2, para aumentar o leque de possíveis palavras a agrupar. Iremos executar esta experiência usando o algoritmo CLARA, e a medida de associação Informação Mútua. Iremos também incluir nas experiências o factor de ponderação, que para esta medida pareceu levar a agrupamentos mais puros.

Como seria de esperar (Anexo C.1), os resultados são necessariamente diferentes dos casos anteriores com a inclusão de todo um conjunto de palavras novas, algumas delas muito raras, e por vezes até erros ortográficos. Quanto aos agrupamentos em sí, os resultados resumem-se a um grupo muito grande e ruidoso, a várias palavras isoladas e a um conjunto de grupos de 2 a 6 elementos que não sendo totalmente inteligíveis, permitem pelo menos compreender a capacidade dos algoritmos de agrupamentos para casos extremos. Há, por exemplo, vários grupos

que parecem conter palavras estrangeiras (japonês, inglês, holandês), e outros grupos que parecem estar relacionados por erros de extracção da formatação da página web de onde provêm. A principal conclusão que se retira desta experiência é a de que, nas condições actuais, o parâmetro existente é de facto necessário e que a sua ausência iria poluir os resultados finais. A sua fixação automática seria de grande utilidade, fazendo uso, por exemplo, do valor da frequência da palavra inicial. A alternativa passaria por tentar incluir um factor de compensação para o efeito de promoção de palavras raras introduzido sistematicamente pela fórmula da Informação Mútua.

É também importante poder verificar o resultado de exigir, através da manipulação do mesmo parâmetro, que as palavras a considerar possuam uma frequência mínima muito elevada, situação oposta à apresentada anteriormente. No Anexo C.2 encontra-se o resultado do agrupamento fixando a frequência mínima das palavras em 20000. Isto é equivalente a reduzir o léxico das palavras co-ocorrente a apenas 8548, ou seja pouco mais de 0,125 % do léxico armazenado no BACO. Por razões de limitação de poder computacional, reduziu-se também o número máximo de palavras a considerar para apenas 200, já que as palavras envolvidas obrigam a indexar em memória muitas mais co-ocorrências. Em conformidade reduziu-se o número de agrupamentos solicitados para apenas 40.

O resultado (Anexo C.2) inclui um conjunto de palavras muito frequentes, cujo agrupamento poderia ser partida potencialmente mais difícil já que a probabilidade da sua co-ocorrência mútua é muito elevada e a matriz de co-ocorrências deverá conter valores muito homogéneos. É quase imediatamente possível verificar que o sentido "vegetal" da palavra "planta" foi quase completamente eliminado dos agrupamentos formados já que estará associado a palavras menos frequentes. Ainda assim, é possivel encontrar alguns agrupamentos inteligíveis, embora com uma relação ainda distante com o sentido "arquitectura" da palavra "planta".

Um outro dado relevante, apesar da comparação não ser directa, é o facto de este ter sido o conjunto de agrupamentos mais homogéneo que obtemos. Há um número relativamente grande de grupos de média ou grande dimensão, quando comparado o resultado com o das outras experiências. Isto poderá eventualmente dever-se à maior homogeneidade da matriz de co-ocorrências, que terá poucos elementos a zero ou de valor reduzido, dada a elevada probabilidade de ocorrência entre todas as palavras.

Como conclusão desta experiência fica a certeza de que as opções de filtragem têm um impacto enorme no resultado final e obrigam a repensar toda esta fase com o máximo de cuidado em trabalhos futuros.

9 Algumas notas sobre os dados

Durante o desenvolvimento do trabalho, e por limitações do espaço em disco foi usada uma tabela de co-ocorrências produzida a partir de uma pequena amostra do WPT03 (menos de 2%) que continha apenas 13 milhões de tuplos. Esta informação, apesar de parecer muito razoável, verificou-se ser insuficiente para a construção sistemática de agrupamentos inteligíveis. Foi interessante verificar que quando corremos a nossa aplicação sobre a totalidade dos dados que tínhamos disponíveis, os resultados começaram a melhorar imenso, sendo esses os que apresentamos em todos os anexos.

Não nos foi possível experimentar determinar qual o ponto, em termos de quantidade de dados, a partir do qual este efeito de emergência ocorre, mas essa seria certamente uma questão muito pertinente. Também não sabemos em que ponto da curva de desempenho nos encontramos neste momento: será que os resultados podem melhorar por simples adição de novos dados ou será que já atingimos o ponto de saturação e necessitamos de usar técnicas mais sofisticadas? É para nós claro que, por agora, aumentar a quantidade de dados não é prioritário mas sim seguir o caminho de melhorar a qualidade do seu processamento. O ponto crítico por passará, por um lado, por sofisticar as medidas de associação usadas, por outro por procurar explorar informação linguística, que neste trabalho foi totalmente ignorada.

10 Conclusões

Neste trabalho procuramos testar as possibilidades de utilização de algoritmos de agrupamento sobre uma base de dados de co-ocorrências. Foram informalmente comparados os resultados de vários algortimos de agrupamento, tendo ficado a sugestão que a utlização de algoritmos de particionamento como o PAM ou o CLARA parecem ser mais apropriados que os algortimos hierarquícos em lidar com este género de dados de grande dimensionalidade. Um ponto crucial, cuja

exploração mais detalhada ficará para trabalho futuro, prende-se a escolha da medida de associação. Nas experiências realizadas, este revelou ser um parâmetro que possui grande impacto no resultado final do agrupamento, quer a nível da produção da matriz de dados a agrupar, quer anteriormente na fase de selecção das palavras a considerar. A escolha apropriada da medida de associação e dos dados base que permitem o seu cálculo parecem assim ser os pontos essenciais a explorar em trabalho futuro.
Como conclusão final, refira-se o facto de as técnicas de agrupamento se terem mostrado promissoras no tratamento da informação relativa às co-ocorrências léxicais sendo por isso de todo o interesse a continuação da sua aplicação a este tipo de problemas, especialmente nos domínios da desambiguação e descoberta de sentidos.

Referências

Berkhin, Pavel. 2002. Survey of Clustering Data Mining Techniques, Accrue Software.

Church, K. & Hanks, P. 1990. Word association norms, mutual information, and lexicography Computational Linguistics, 1990, 16(1), 22–29

Evert, Stefan. 2005. The statistics of word cooccurrences : word pairs and collocations. PhD thesis. Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2005.

Gomes, D.; Cardoso, N.; Seco, N.; Santos, D. & Silva, M. WPT 03: Portuguese Web at your fingertips Submitted, 2006.

Schone, P. & Jurafsky, D. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? Computational Linguistics, Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, 2001, pp 100–108

Yarowsky, David. 1993. One Sense Per Collocation. Proceedings of the ARPA Human Language Technology Workshop.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. 1995. pp 189–196

Anexo A.1
Agrupamentos usado algoritmo DIANA

Cluster 0: topográfico
Cluster 1: quintãs porteira sacadas quintã chafarizes emcel judaicas renascentistas sócio-económicos celorico balcões mogadouro
Cluster 2: àescala zonamento condicionantes justificativa descritiva pretensão
Cluster 3: mesquitela rapa ratoeira prados fornotelheiro
Cluster 4: cadastral topográfica
Cluster 5: tubérculos caules vegetativo míldio susceptibilidade floração polpa batateira
Cluster 6: ciconia
Cluster 7: octogonal abóbada cúpula capela-mor sacristia claustro
Cluster 8: trapalhão subscritos miróbriga cannabis fotografado zenith cala coca exigente rodar cresce filas imaginário prestador lema
Cluster 9: alçados cérceas
Cluster 10: complementação àfundação camionagem continha gare quiosque
Cluster 11: vãos
Cluster 12: caule pigmentação corola espiga limbo
Cluster 13: culinário medianamente nemátodos grânulos enrolamento parasita imune parasitas rama proteínas
Cluster 14: polidesportivo
Cluster 15: quadrangular rectangular quadrada pilares
Cluster 16: invasoras espécimes indígenas detenham originária deter detenção
Cluster 17: longitudinais
Cluster 18: parcelar
Cluster 19: alçado
Cluster 20: velosa lageosa calvário soeiro linhares
Cluster 21: aglomerado levantamento implantação traçado envolvente arranjo arruamentos esgotos pormenor urbano localização aglomerados enquadramento parcela drenagem especialidades edificado terreno urbanas urbana limites prédio salvaguarda edifícios urbanização recuperação espaços existente reconstrução urbanos construções abastecimento ocupação imóvel zona eléctrica destinada tipologia vias exteriores agrícola cobertura edifício terrenos águas reabilitação conservação plano área ribeira substituição áreas elementos histórico Áreas lote
Cluster 22: fachadas pisos fachada poente piso nascente cave coberta muros pavimento construída
Cluster 23: poligonal
Cluster 24: medicinal
Cluster 25: falésia arade varanda banho castanho quintal
Cluster 26: esclarecendo extractos vigente exista indicando extracto
Cluster 27: plant
Cluster 28: anexa escala síntese ecológica indicação adjacentes estimativa devidamente específico regulamentares existir abrangendo situa operação construídos anexos cedência alvará elaborada insere administrativas trate arranjos restrições existam fracção referentes utilidade requerente futura reserva cadastro aditamento complementares aplicáveis turística predial destinadas regulamento taxas objecto prévia
Cluster 29: topográficos levantamentos
Cluster 30: agricultor
Cluster 31: expropriações àcâmara
Cluster 32: longitudinal
Cluster 33: servidões ordenamento volumetria cércea edificações delimitação loteamento urbanística urbanísticos arruamento viária estacionamentos perímetro lotes condicionamentos parcelas demolição dimensionamento habitacional edificação usos infra-estruturas
Cluster 34: soleira desenhadas
Cluster 35: urbanizável urbanizáveis contíguos delimitada
Cluster 36: espaços-canais operativa
Cluster 37: assinalando preenchida calendarização subscrito ratificado actualizada estatísticos entenda cópia curvas junção comunicar notificação apoiada terço optimização estiver precisa
Cluster 38: operativas
Cluster 39: telheiro
Cluster 40: arcadas
Cluster 41: toponímia molduras decorativos numeração notáveis aldeias janelas religiosos pormenores beira
Cluster 42: anexas
Cluster 43: alvenaria
Cluster 44: ornamental ecologia
Cluster 45: perímetros paisagístico
Cluster 46: pluviais transversais perfis arquitectónica topografia pavimentos
Cluster 47: fossas escoamento
Cluster 48: telhados nave capelas naves calçada
Cluster 49: cotas alinhamento variante cortes histórica circular cota eixo
Cluster 50: semente sementeira variedades insectos estufa sementes
Cluster 51: cotada
Cluster 52: àentrada altitude irregular
Cluster 53: plantas solo folha profundidade
Cluster 54: lusófona martelo mediateca paralelos font perna espólio papelaria vende
Cluster 55: primárias
Cluster 56: arbustos
Cluster 57: organigrama
Cluster 58: perpendicular
Cluster 59: moinhos mondego atalaia viçosa chão mangualde
Cluster 60: telhado muralha telha moinho garagem
Cluster 61: designa incorporados estupefacientes
Cluster 62: escassa abundante amarela lisa porte
Cluster 63: rústico
Cluster 64: definindo designadas impliquem destinam máximos predominantemente
Cluster 65: plantação solos espécies vegetal florestais
Cluster 66: susceptível integrante licenciadas
Cluster 67: cultivo
Cluster 68: grão seca folhas raiz fresco raízes tronco fruto milho
Cluster 69: povoado muralhas
Cluster 70: sanitárias
Cluster 71: superficiais
Cluster 72: armazenagem aterro efluentes
Cluster 73: vegetação escadas rochas
Cluster 74: transversal

---

9 clusters com 6 elementos!
1 clusters com 11 elementos!
5 clusters com 3 elementos!
1 clusters com 7 elementos!
2 clusters com 9 elementos!
10 clusters com 2 elementos!
1 clusters com 12 elementos!
1 clusters com 15 elementos!
1 clusters com 22 elementos!
1 clusters com 42 elementos!
2 clusters com 8 elementos!
27 clusters com 1 elementos!
3 clusters com 4 elementos!
1 clusters com 18 elementos!
1 clusters com 56 elementos!
2 clusters com 10 elementos!
7 clusters com 5 elementos!

Anexo A.2
Agrupamentos usado algoritmo AGNES

Cluster 0: topográfico
Cluster 1: quintãs porteira sacadas quintã chafarizes emcel judaicas renascentistas molduras balcões
Cluster 2: àescala condicionantes aglomerado fachadas levantamento justificativa anexa servidões ordenamento implantação escala desenhadas edificações delimitação descritiva perímetros síntese pisos traçado cotas loteamento urbanística urbanísticos plantas viária fachada envolvente estacionamentos perímetro ecológica exista indicando lotes arruamentos esgotos alinhamento pormenor pretensão indicação condicionamentos urbano localização aglomerados enquadramento devidamente solos parcela drenagem específico regulamentares poente existir edificado abrangendo situa terreno urbanas urbana piso operação parcelas limites integrante demolição prédio construídos anexos salvaguarda espécies edifícios cedência habitacional urbanização recuperação nascente espaços existente alvará edificação cortes reconstrução urbanos administrativas construções trate abastecimento solo arranjos restrições ocupação existam fracção coberta referentes utilidade requerente imóvel muros histórica zona usos infra-estruturas futura reserva destinada aditamento circular tipologia vias complementares aplicáveis exteriores turística pavimento vegetal agrícola cota predial cobertura edifício terrenos águas reabilitação destinadas conservação plano regulamento área construída taxas substituição objecto eixo prévia áreas elementos florestais Áreas precisa lote
Cluster 3: mesquitela rapa ratoeira fornotelheiro velosa lageosa calvário gare soeiro linhares viçosa mangualde
Cluster 4: cadastral cadastro
Cluster 5: tubérculos
Cluster 6: ciconia
Cluster 7: octogonal quadrangular rectangular quadrada abóbada cúpula capela-mor sacristia claustro
Cluster 8: trapalhão culinário miróbriga cannabis fotografado cala imune coca lusófona martelo exigente font rodar perna cresce filas imaginário vende lema
Cluster 9: prados
Cluster 10: alçados
Cluster 11: complementação àfundação zenith continha paralelos quiosque
Cluster 12: vãos decorativos arquitectónica janelas
Cluster 13: caule pigmentação caules vegetativo floração
Cluster 14: polidesportivo
Cluster 15: zonamento
Cluster 16: corola espiga limbo
Cluster 17: subscritos preenchida calendarização subscrito estimativa incorporados ratificado actualizada estatísticos entenda elaborada cópia insere curvas junção comunicar notificação apoiada prestador terço optimização estiver
Cluster 18: cérceas
Cluster 19: invasoras espécimes indígenas
Cluster 20: longitudinais
Cluster 21: parcelar
Cluster 22: alçado
Cluster 23: topográfica
Cluster 24: medianamente nemátodos míldio grânulos enrolamento polpa parasita batateira parasitas insectos proteínas
Cluster 25: poligonal
Cluster 26: sócio-económicos toponímia
Cluster 27: medicinal
Cluster 28: falésia arade
Cluster 29: esclarecendo extractos assinalando vigente extracto
Cluster 30: plant
Cluster 31: topográficos levantamentos
Cluster 32: agricultor
Cluster 33: expropriações
Cluster 34: longitudinal transversal
Cluster 35: soleira volumetria cércea arruamento
Cluster 36: urbanizável urbanizáveis contíguos delimitada
Cluster 37: espaços-canais
Cluster 38: operativas operativa
Cluster 39: telheiro
Cluster 40: arcadas
Cluster 41: susceptibilidade
Cluster 42: anexas
Cluster 43: celorico pluviais mogadouro transversais arranjo perfis moinhos numeração àcâmara topografia pavimentos notáveis especialidades mondego aldeias variante religiosos pormenores chão beira eléctrica ribeira histórico
Cluster 44: camionagem àentrada
Cluster 45: alvenaria sanitárias escadas
Cluster 46: ornamental
Cluster 47: fossas
Cluster 48: telhados pilares
Cluster 49: detenham designa originária estupefacientes deter detenção
Cluster 50: semente sementeira grão variedades sementes milho
Cluster 51: cotada
Cluster 52: primárias
Cluster 53: arbustos
Cluster 54: organigrama mediateca espólio papelaria
Cluster 55: perpendicular
Cluster 56: telhado telha moinho calçada
Cluster 57: escassa
Cluster 58: adjacentes definindo designadas impliquem dimensionamento armazenagem escoamento destinam máximos aterro efluentes predominantemente
Cluster 59: rústico
Cluster 60: plantação
Cluster 61: nave capelas naves
Cluster 62: susceptível licenciadas
Cluster 63: cultivo
Cluster 64: paisagístico
Cluster 65: altitude irregular
Cluster 66: povoado muralha muralhas
Cluster 67: seca folhas porte folha raiz raízes fruto profundidade
Cluster 68: superficiais
Cluster 69: abundante rama amarela lisa fresco
Cluster 70: atalaia
Cluster 71: ecologia estufa
Cluster 72: cave garagem
Cluster 73: vegetação rochas
Cluster 74: varanda banho tronco castanho quintal

---

3 clusters com 6 elementos!
1 clusters com 11 elementos!
5 clusters com 3 elementos!
1 clusters com 9 elementos!
13 clusters com 2 elementos!
2 clusters com 12 elementos!
1 clusters com 8 elementos!
1 clusters com 22 elementos!
5 clusters com 4 elementos!
35 clusters com 1 elementos!
1 clusters com 23 elementos!
1 clusters com 19 elementos!
1 clusters com 10 elementos!
1 clusters com 140 elementos!
4 clusters com 5 elementos!

Anexo A.3
Agrupamentos usado algoritmo PAM

Cluster 0: topográfico
Cluster 1: quintãs porteira sacadas chafarizes emcel judaicas renascentistas sócio-económicos
Cluster 2: àescala justificativa desenhadas descritiva exista indicando indicação extracto devidamente existir abrangendo situa sanitárias trate existam complementares taxas objecto prévia
Cluster 3: mesquitela rapa ratoeira fornotelheiro velosa
Cluster 4: cadastral
Cluster 5: tubérculos caule pigmentação caules vegetativo floração polpa
Cluster 6: ciconia
Cluster 7: quintã celorico mogadouro beira
Cluster 8: octogonal quadrangular abóbada capela-mor sacristia claustro
Cluster 9: trapalhão culinário subscritos miróbriga cannabis fotografado grânulos zenith cala imune coca lema
Cluster 10: prados
Cluster 11: alçados
Cluster 12: complementação àfundação continha quiosque
Cluster 13: vãos
Cluster 14: polidesportivo
Cluster 15: zonamento
Cluster 16: corola limbo
Cluster 17: condicionantes aglomerado levantamento ordenamento edificações delimitação perímetros síntese viária envolvente perímetro ecológica condicionamentos adjacentes topografia aglomerados solos paisagístico edificado urbanas construídos dimensionamento salvaguarda espécies habitacional urbanos solo arranjos usos infra-estruturas efluentes turística predominantemente vegetal destinadas regulamento áreas profundidade florestais Áreas
Cluster 18: cérceas
Cluster 19: invasoras espécimes indígenas
Cluster 20: longitudinais transversais perfis
Cluster 21: parcelar
Cluster 22: alçado
Cluster 23: rectangular quadrada cúpula
Cluster 24: topográfica
Cluster 25: medianamente nemátodos míldio enrolamento susceptibilidade batateira rama
Cluster 26: fachadas fachada telhado arquitectónica rústico poente janelas piso edifícios coberta imóvel tipologia pavimento edifício construída
Cluster 27: poligonal
Cluster 28: espiga
Cluster 29: medicinal
Cluster 30: falésia arade
Cluster 31: esclarecendo extractos assinalando
Cluster 32: plant parasita parasitas filas insectos proteínas
Cluster 33: anexa implantação escala pisos traçado plantas lotes pormenor urbano localização enquadramento parcela especialidades terreno urbana operação parcelas limites demolição prédio anexos cedência urbanização recuperação nascente espaços existente edificação cortes reconstrução construções ocupação muros zona reserva eléctrica destinada vias exteriores agrícola cobertura terrenos águas reabilitação conservação plano área substituição eixo elementos lote
Cluster 34: topográficos levantamentos
Cluster 35: agricultor
Cluster 36: expropriações anexas camionagem àentrada escoamento elaborada referentes futura circular apoiada
Cluster 37: longitudinal transversal
Cluster 38: servidões vigente definindo estimativa susceptível específico regulamentares designadas impliquem integrante licenciadas armazenagem destinam máximos administrativas restrições utilidade cadastro aplicáveis
Cluster 39: soleira cotas estacionamentos alinhamento cave aterro cota
Cluster 40: urbanizável urbanizáveis contíguos
Cluster 41: espaços-canais
Cluster 42: operativas operativa
Cluster 43: telheiro
Cluster 44: lageosa calvário soeiro linhares
Cluster 45: preenchida calendarização designa subscrito incorporados actualizada entenda estupefacientes cópia insere irregular junção comunicar notificação terço estiver precisa
Cluster 46: volumetria cércea
Cluster 47: arcadas
Cluster 48: toponímia arranjo primárias decorativos numeração àcâmara notáveis aldeias variante religiosos pormenores histórica histórico
Cluster 49: molduras balcões
Cluster 50: alvenaria
Cluster 51: ornamental
Cluster 52: pluviais arruamentos esgotos pavimentos drenagem abastecimento
Cluster 53: fossas
Cluster 54: telhados altitude telha escadas varanda banho pilares castanho papelaria quintal garagem
Cluster 55: detenham originária deter detenção
Cluster 56: semente sementeira cultivo variedades sementes milho
Cluster 57: loteamento urbanística urbanísticos arruamento pretensão alvará fracção requerente aditamento predial
Cluster 58: cotada
Cluster 59: gare atalaia viçosa mangualde
Cluster 60: lusófona ratificado exigente mediateca font cresce espólio imaginário prestador vende
Cluster 61: delimitada
Cluster 62: arbustos
Cluster 63: organigrama estatísticos ecologia estufa curvas optimização
Cluster 64: perpendicular
Cluster 65: martelo paralelos rodar perna
Cluster 66: moinhos mondego chão ribeira moinho calçada
Cluster 67: escassa
Cluster 68: plantação
Cluster 69: nave capelas naves
Cluster 70: grão seca folhas porte folha raiz raízes tronco fruto
Cluster 71: povoado muralha muralhas
Cluster 72: superficiais
Cluster 73: abundante amarela lisa fresco
Cluster 74: vegetação rochas

---

1 clusters com 11 elementos!
3 clusters com 7 elementos!
1 clusters com 17 elementos!
8 clusters com 2 elementos!
29 clusters com 1 elementos!
1 clusters com 13 elementos!
6 clusters com 6 elementos!
1 clusters com 40 elementos!
7 clusters com 3 elementos!
1 clusters com 9 elementos!
1 clusters com 51 elementos!
1 clusters com 12 elementos!
1 clusters com 15 elementos!
1 clusters com 8 elementos!
7 clusters com 4 elementos!
2 clusters com 19 elementos!
3 clusters com 10 elementos!
1 clusters com 5 elementos!

Anexo A.4
Agrupamentos usado algoritmo CLARA

Cluster 0: topográfico longitudinais topográficos transversais perfis levantamentos
Cluster 1: quintãs porteira sacadas quintã chafarizes emcel judaicas renascentistas sócio-económicos
Cluster 2: àescala condicionantes justificativa esclarecendo extractos servidões desenhadas descritiva vigente exista indicando pretensão indicação extracto definindo estimativa devidamente susceptível específico regulamentares existir abrangendo situa impliquem integrante licenciadas sanitárias armazenagem máximos administrativas trate restrições existam fracção utilidade requerente aditamento complementares aplicáveis predial taxas objecto prévia
Cluster 3: mesquitela rapa ratoeira fornotelheiro velosa
Cluster 4: cadastral
Cluster 5: tubérculos
Cluster 6: ciconia
Cluster 7: octogonal
Cluster 8: trapalhão complementação culinário corola subscritos miróbriga nemátodos cannabis fotografado grânulos enrolamento zenith cala parasita imune coca lusófona exigente rodar imaginário prestador lema proteínas
Cluster 9: prados
Cluster 10: alçados
Cluster 11: vãos
Cluster 12: caule pigmentação medianamente caules limbo vegetativo míldio susceptibilidade floração polpa parasitas
Cluster 13: polidesportivo
Cluster 14: quadrangular quadrada abóbada cúpula capela-mor sacristia claustro naves
Cluster 15: zonamento
Cluster 16: àfundação camionagem continha paralelos quiosque
Cluster 17: cérceas
Cluster 18: invasoras espécimes indígenas
Cluster 19: parcelar
Cluster 20: alçado
Cluster 21: rectangular pilares
Cluster 22: topográfica expropriações anexas designadas escoamento elaborada referentes futura cadastro circular apoiada optimização
Cluster 23: aglomerado levantamento anexa ordenamento implantação escala edificações delimitação síntese pisos traçado cotas loteamento urbanística arruamento plantas viária envolvente perímetro ecológica lotes alinhamento pormenor adjacentes urbano localização topografia rústico aglomerados enquadramento solos parcela especialidades poente paisagístico edificado terreno urbanas urbana piso operação parcelas limites demolição prédio construídos anexos dimensionamento salvaguarda espécies edifícios cedência habitacional urbanização recuperação nascente espaços existente alvará edificação cortes reconstrução urbanos cave aterro construções solo arranjos ocupação coberta imóvel muros zona usos infra-estruturas reserva eléctrica destinada tipologia vias efluentes exteriores turística predominantemente vegetal agrícola cota cobertura edifício terrenos águas reabilitação destinadas conservação plano regulamento área substituição eixo áreas profundidade elementos florestais Áreas lote
Cluster 24: fachadas fachada escadas pavimento construída
Cluster 25: poligonal
Cluster 26: espiga plant
Cluster 27: medicinal rama
Cluster 28: falésia
Cluster 29: agricultor
Cluster 30: longitudinal transversal
Cluster 31: soleira volumetria cércea contíguos estacionamentos
Cluster 32: urbanizável perímetros urbanizáveis urbanísticos condicionamentos
Cluster 33: espaços-canais
Cluster 34: assinalando preenchida calendarização designa subscrito incorporados ratificado actualizada estatísticos entenda cópia destinam filas insere irregular curvas junção comunicar notificação terço estiver precisa
Cluster 35: arade
Cluster 36: operativas
Cluster 37: telheiro
Cluster 38: lageosa calvário gare soeiro linhares atalaia viçosa mangualde
Cluster 39: arcadas
Cluster 40: toponímia arranjo moinhos decorativos arquitectónica numeração àcâmara notáveis mondego aldeias janelas variante religiosos pormenores chão capelas histórica ribeira histórico
Cluster 41: molduras balcões
Cluster 42: celorico mogadouro beira
Cluster 43: alvenaria
Cluster 44: ornamental
Cluster 45: pluviais arruamentos esgotos pavimentos drenagem abastecimento
Cluster 46: fossas
Cluster 47: telhados
Cluster 48: detenham
Cluster 49: semente plantação cultivo variedades insectos sementes
Cluster 50: cotada
Cluster 51: àentrada
Cluster 52: operativa
Cluster 53: batateira sementeira
Cluster 54: primárias
Cluster 55: delimitada
Cluster 56: arbustos nave altitude varanda banho papelaria quintal garagem calçada
Cluster 57: organigrama
Cluster 58: perpendicular
Cluster 59: martelo
Cluster 60: telhado
Cluster 61: escassa
Cluster 62: grão milho
Cluster 63: mediateca font cresce espólio vende
Cluster 64: originária estupefacientes deter detenção
Cluster 65: povoado muralha muralhas
Cluster 66: seca folhas porte folha raiz raízes tronco fruto
Cluster 67: perna
Cluster 68: superficiais
Cluster 69: abundante amarela lisa fresco
Cluster 70: telha
Cluster 71: ecologia estufa
Cluster 72: vegetação rochas
Cluster 73: castanho
Cluster 74: moinho

---

3 clusters com 6 elementos!
2 clusters com 11 elementos!
3 clusters com 3 elementos!
2 clusters com 9 elementos!
1 clusters com 12 elementos!
9 clusters com 2 elementos!
1 clusters com 22 elementos!
3 clusters com 8 elementos!
2 clusters com 4 elementos!
39 clusters com 1 elementos!
1 clusters com 19 elementos!
1 clusters com 23 elementos!
1 clusters com 43 elementos!
1 clusters com 105 elementos!
6 clusters com 5 elementos!

Anexo A.5
Agrupamentos usado algoritmo FANNY

Cluster 0: topográfico àescala cadastral octogonal prados alçados vãos culinário polidesportivo quadrangular zonamento corola subscritos cérceas longitudinais parcelar alçado rectangular topográfica velosa aglomerado fachadas espécimes poligonal sócio-económicos medicinal levantamento justificativa esclarecendo limbo cannabis plant anexa topográficos agricultor expropriações extractos longitudinal servidões urbanizável ordenamento implantação espaços-canais quadrada assinalando arade operativas escala telheiro lageosa cúpula volumetria desenhadas arcadas edificações toponímia susceptibilidade anexas capela-mor cala molduras camionagem alvenaria descritiva balcões parasita ornamental síntese vigente calvário imune pluviais transversais coca fossas urbanizáveis traçado telhados contíguos detenham semente continha urbanística cotada gare àentrada urbanísticos arruamento plantas fachada envolvente lusófona calendarização perímetro ecológica arranjo primárias arbustos organigrama exista soeiro perfis martelo levantamentos moinhos indicando decorativos arquitectónica linhares lotes arruamentos claustro numeração esgotos alinhamento pormenor pretensão indicação àcâmara extracto condicionamentos adjacentes urbano localização subscrito topografia sementeira definindo aglomerados plantação estimativa notáveis nave enquadramento devidamente solos incorporados susceptível cultivo parcela drenagem especialidades ratificado actualizada específico regulamentares exigente poente mondego existir aldeias edificado abrangendo mediateca designadas situa janelas terreno paralelos urbanas urbana estatísticos piso entenda impliquem operação altitude parcelas limites integrante demolição font seca estupefacientes prédio construídos licenciadas sanitárias anexos dimensionamento salvaguarda espécies edifícios rodar cedência variante habitacional urbanização muralha perna armazenagem recuperação abundante escoamento religiosos espaços quiosque rama existente alvará elaborada edificação telha cortes cópia cresce reconstrução destinam muralhas amarela lisa máximos ecologia variedades filas folhas urbanos cave pormenores insere administrativas estufa chão porte aterro construções trate capelas abastecimento solo arranjos restrições deter ocupação existam beira fracção vegetação coberta folha referentes utilidade requerente transversal imóvel irregular escadas muros sementes curvas junção histórica zona varanda usos infra-estruturas futura reserva eléctrica raiz fresco imaginário destinada cadastro aditamento banho circular comunicar tipologia vias complementares pilares efluentes raízes aplicáveis exteriores turística predominantemente pavimento rochas vegetal agrícola cota predial cobertura edifício terrenos águas notificação reabilitação apoiada destinadas conservação plano tronco prestador regulamento papelaria área ribeira terço optimização construída fruto taxas substituição objecto eixo prévia áreas detenção vende quintal profundidade milho lema proteínas moinho garagem elementos histórico florestais estiver calçada Áreas precisa lote
Cluster 1: quintãs quintã emcel renascentistas
Cluster 2: mesquitela rapa ratoeira fornotelheiro
Cluster 3: porteira
Cluster 4: tubérculos caules floração
Cluster 5: sacadas chafarizes judaicas
Cluster 6: ciconia
Cluster 7: trapalhão complementação àfundação fotografado zenith
Cluster 8: caule pigmentação
Cluster 9: condicionantes
Cluster 10: invasoras indígenas
Cluster 11: miróbriga
Cluster 12: medianamente míldio enrolamento polpa parasitas
Cluster 13: nemátodos grânulos
Cluster 14: espiga
Cluster 15: falésia
Cluster 16: vegetativo batateira
Cluster 17: soleira
Cluster 18: abóbada
Cluster 19: preenchida
Cluster 20: cércea
Cluster 21: celorico mogadouro
Cluster 22: delimitação
Cluster 23: perímetros
Cluster 24: pisos
Cluster 25: sacristia
Cluster 26: cotas
Cluster 27: loteamento
Cluster 28: operativa
Cluster 29: viária
Cluster 30: estacionamentos
Cluster 31: delimitada
Cluster 32: perpendicular
Cluster 33: telhado
Cluster 34: designa
Cluster 35: escassa
Cluster 36: rústico
Cluster 37: pavimentos
Cluster 38: grão
Cluster 39: paisagístico
Cluster 40: originária
Cluster 41: povoado
Cluster 42: superficiais
Cluster 43: nascente
Cluster 44: atalaia
Cluster 45: espólio
Cluster 46: insectos
Cluster 47: viçosa
Cluster 48: mangualde
Cluster 49: naves
Cluster 50: castanho
Cluster 51:
Cluster 52:
Cluster 53:
Cluster 54:
Cluster 55:
Cluster 56:
Cluster 57:
Cluster 58:
Cluster 59:
Cluster 60:
Cluster 61:
Cluster 62:
Cluster 63:
Cluster 64:
Cluster 65:
Cluster 66:
Cluster 67:
Cluster 68:
Cluster 69:
Cluster 70:
Cluster 71:
Cluster 72:
Cluster 73:
Cluster 74:

---

5 clusters com 2 elementos!
1 clusters com 327 elementos!
39 clusters com 1 elementos!
24 clusters com 0 elementos!
2 clusters com 3 elementos!
2 clusters com 4 elementos!
2 clusters com 5 elementos!

Anexo B.1
Agrupamentos usado algoritmo CLARA, PMI e Ponderação

Cluster 0: topográfico
Cluster 1: quintãs porteira
Cluster 2: àescala condicionantes
Cluster 3: mesquitela rapa ratoeira fornotelheiro velosa
Cluster 4: cadastral
Cluster 5: tubérculos
Cluster 6: sacadas
Cluster 7: ciconia
Cluster 8: quintã chafarizes emcel judaicas renascentistas
Cluster 9: octogonal
Cluster 10: trapalhão miróbriga cannabis fotografado grânulos cala parasita imune coca lusófona exigente mediateca font estupefacientes rodar rama cresce filas deter imaginário prestador vende lema proteínas
Cluster 11: prados
Cluster 12: alçados
Cluster 13: complementação àfundação
Cluster 14: vãos
Cluster 15: caule pigmentação caules vegetativo floração polpa
Cluster 16: culinário medianamente
Cluster 17: polidesportivo
Cluster 18: quadrangular
Cluster 19: zonamento
Cluster 20: corola limbo
Cluster 21: subscritos assinalando preenchida cotada calendarização organigrama exista indicando designa extracto subscrito estimativa incorporados ratificado actualizada situa estatísticos entenda elaborada cópia espólio insere transversal irregular curvas junção aditamento comunicar pilares notificação apoiada terço optimização estiver precisa
Cluster 22: cérceas
Cluster 23: invasoras indígenas
Cluster 24: longitudinais
Cluster 25: parcelar
Cluster 26: alçado
Cluster 27: rectangular
Cluster 28: topográfica
Cluster 29: nemátodos míldio enrolamento batateira
Cluster 30: aglomerado
Cluster 31: fachadas levantamento topográficos implantação pluviais transversais traçado plantas envolvente perfis arquitectónica arruamentos esgotos pormenor àcâmara urbano localização rústico pavimentos enquadramento drenagem salvaguarda edifícios abastecimento ocupação tipologia
Cluster 32: espécimes
Cluster 33: poligonal
Cluster 34: espiga
Cluster 35: sócio-económicos
Cluster 36: medicinal
Cluster 37: justificativa desenhadas
Cluster 38: falésia
Cluster 39: esclarecendo extractos
Cluster 40: plant
Cluster 41: anexa ordenamento escala anexas delimitação descritiva síntese pisos cotas loteamento arruamento viária fachada perímetro ecológica arranjo primárias delimitada levantamentos lotes alinhamento pretensão indicação adjacentes topografia aglomerados devidamente solos parcela especialidades poente paisagístico edificado abrangendo terreno urbanas urbana piso operação parcelas limites demolição prédio construídos anexos espécies cedência variante habitacional urbanização recuperação escoamento nascente espaços existente alvará edificação cortes reconstrução urbanos cave aterro construções solo arranjos fracção vegetação coberta folha referentes requerente imóvel muros histórica zona usos infra-estruturas futura reserva eléctrica destinada circular vias exteriores turística pavimento rochas vegetal agrícola cota predial cobertura edifício terrenos águas reabilitação conservação plano regulamento área construída taxas substituição eixo prévia áreas profundidade elementos histórico florestais Áreas lote
Cluster 42: agricultor
Cluster 43: expropriações
Cluster 44: longitudinal
Cluster 45: servidões urbanizável volumetria cércea edificações urbanizáveis contíguos urbanística urbanísticos estacionamentos condicionamentos
Cluster 46: soleira
Cluster 47: espaços-canais
Cluster 48: quadrada
Cluster 49: arade
Cluster 50: abóbada capela-mor sacristia claustro
Cluster 51: operativas
Cluster 52: telheiro
Cluster 53: lageosa balcões calvário soeiro moinhos decorativos linhares numeração notáveis mondego aldeias janelas religiosos pormenores chão beira mangualde ribeira
Cluster 54: cúpula
Cluster 55: arcadas
Cluster 56: toponímia
Cluster 57: susceptibilidade
Cluster 58: zenith camionagem continha gare àentrada paralelos quiosque viçosa
Cluster 59: molduras
Cluster 60: celorico mogadouro
Cluster 61: alvenaria
Cluster 62: ornamental
Cluster 63: perímetros
Cluster 64: vigente detenham operativa definindo susceptível específico regulamentares existir designadas impliquem originária integrante licenciadas sanitárias dimensionamento armazenagem destinam máximos administrativas trate restrições existam utilidade cadastro complementares efluentes aplicáveis predominantemente destinadas objecto detenção
Cluster 65: fossas
Cluster 66: telhados
Cluster 67: semente sementeira plantação grão seca folhas estufa porte raiz raízes fruto milho
Cluster 68: arbustos
Cluster 69: perpendicular
Cluster 70: martelo nave altitude povoado muralha perna telha atalaia muralhas amarela capelas naves escadas varanda fresco banho tronco castanho papelaria quintal moinho garagem calçada
Cluster 71: telhado
Cluster 72: parasitas cultivo abundante lisa ecologia variedades insectos sementes
Cluster 73: escassa
Cluster 74: superficiais

---

1 clusters com 35 elementos!
1 clusters com 6 elementos!
2 clusters com 11 elementos!
1 clusters com 26 elementos!
1 clusters com 12 elementos!
9 clusters com 2 elementos!
2 clusters com 8 elementos!
1 clusters com 112 elementos!
49 clusters com 1 elementos!
2 clusters com 4 elementos!
1 clusters com 18 elementos!
1 clusters com 24 elementos!
1 clusters com 23 elementos!
1 clusters com 31 elementos!
2 clusters com 5 elementos!

Anexo B.2
Agrupamentos usado algoritmo CLARA, Frequência, sem Ponderação, com Normalização

Cluster 0: topográfico quintãs porteira quintã vãos judaicas cérceas renascentistas sócio-económicos toponímia molduras balcões transversais fachada levantamentos decorativos arquitectónica arruamentos numeração àcâmara pavimentos notáveis aldeias janelas salvaguarda religiosos pormenores capelas ocupação varanda tipologia histórico
Cluster 1: àescala cadastral zonamento subscritos invasoras longitudinais rectangular aglomerado fachadas justificativa esclarecendo anexa expropriações longitudinal servidões soleira urbanizável quadrada assinalando volumetria desenhadas cércea edificações alvenaria delimitação perímetros vigente pluviais urbanizáveis traçado cotas urbanística urbanísticos arruamento viária envolvente perímetro ecológica arranjo delimitada exista perpendicular telhado indicando lotes esgotos alinhamento extracto condicionamentos topografia rústico aglomerados plantação estimativa cultivo especialidades específico regulamentares paisagístico edificado abrangendo designadas urbanas estatísticos impliquem parcelas integrante demolição seca construídos sanitárias anexos dimensionamento cedência habitacional armazenagem escoamento alvará telha cortes reconstrução máximos ecologia variedades filas cave administrativas aterro trate arranjos restrições deter existam fracção coberta referentes utilidade transversal irregular escadas muros curvas histórica usos futura eléctrica raiz destinada cadastro aditamento circular comunicar complementares pilares efluentes aplicáveis exteriores turística predominantemente pavimento vegetal cota destinadas optimização construída substituição detenção profundidade garagem florestais estiver
Cluster 2: mesquitela rapa ratoeira prados emcel fornotelheiro lageosa mogadouro soeiro moinhos linhares mondego mangualde naves
Cluster 3: tubérculos medianamente caules vegetativo míldio susceptibilidade floração polpa parasitas escassa susceptível superficiais abundante amarela lisa porte sementes fresco fruto
Cluster 4: sacadas octogonal trapalhão chafarizes alçados complementação culinário polidesportivo quadrangular àfundação corola parcelar alçado topográfica miróbriga velosa poligonal medicinal plant fotografado topográficos agricultor espaços-canais arade abóbada operativas telheiro preenchida cúpula arcadas enrolamento zenith anexas capela-mor cala camionagem parasita ornamental calvário imune sacristia fossas telhados contíguos detenham continha cotada gare àentrada operativa lusófona calendarização estacionamentos primárias arbustos organigrama claustro adjacentes subscrito sementeira definindo incorporados grão ratificado actualizada exigente mediateca situa paralelos entenda altitude originária povoado licenciadas rodar muralha quiosque rama elaborada cresce destinam atalaia muralhas espólio insectos insere viçosa junção imaginário apoiada tronco prestador castanho papelaria terço vende quintal lema proteínas calçada
Cluster 5: ciconia vegetação
Cluster 6: caule pigmentação espiga limbo
Cluster 7: condicionantes extractos descritiva pormenor indicação enquadramento drenagem existir existente edificação cópia notificação prévia precisa
Cluster 8: nemátodos grânulos estufa
Cluster 9: espécimes indígenas espécies
Cluster 10: levantamento
Cluster 11: falésia
Cluster 12: cannabis coca designa estupefacientes
Cluster 13: ordenamento plano
Cluster 14: implantação
Cluster 15: escala
Cluster 16: celorico
Cluster 17: síntese
Cluster 18: pisos
Cluster 19: semente batateira
Cluster 20: loteamento
Cluster 21: plantas
Cluster 22: perfis
Cluster 23: martelo perna
Cluster 24: pretensão requerente zona
Cluster 25: urbano
Cluster 26: localização
Cluster 27: nave cobertura
Cluster 28: devidamente
Cluster 29: solos
Cluster 30: parcela poente nascente predial
Cluster 31: terreno
Cluster 32: urbana
Cluster 33: piso
Cluster 34: operação elementos
Cluster 35: limites
Cluster 36: font
Cluster 37: prédio
Cluster 38: edifícios
Cluster 39: variante
Cluster 40: urbanização
Cluster 41: recuperação
Cluster 42: espaços
Cluster 43: folhas
Cluster 44: urbanos
Cluster 45: chão
Cluster 46: construções
Cluster 47: abastecimento
Cluster 48: solo
Cluster 49: beira
Cluster 50: folha
Cluster 51: imóvel
Cluster 52: infra-estruturas
Cluster 53: reserva
Cluster 54: banho
Cluster 55: vias
Cluster 56: raízes
Cluster 57: rochas
Cluster 58: agrícola
Cluster 59: edifício
Cluster 60: terrenos
Cluster 61: águas
Cluster 62: reabilitação
Cluster 63: conservação
Cluster 64: regulamento
Cluster 65: área
Cluster 66: ribeira
Cluster 67: taxas
Cluster 68: objecto
Cluster 69: eixo
Cluster 70: áreas
Cluster 71: milho
Cluster 72: moinho
Cluster 73: Áreas
Cluster 74: lote

---

1 clusters com 11 elementos!
1 clusters com 131 elementos!
1 clusters com 32 elementos!
3 clusters com 3 elementos!
6 clusters com 2 elementos!
2 clusters com 14 elementos!
3 clusters com 4 elementos!
56 clusters com 1 elementos!
1 clusters com 19 elementos!
1 clusters com 100 elementos!

Anexo B.3
Agrupamentos usado algoritmo CLARA, Frequência, com Ponderação, com normalização

Cluster 0: topográfico levantamento
Cluster 1: quintãs porteira quintã prados alçados vãos emcel cérceas renascentistas fachadas sócio-económicos molduras balcões mogadouro traçado urbanística fachada envolvente moinhos decorativos arquitectónica arruamentos numeração pavimentos notáveis nave mondego aldeias janelas salvaguarda reconstrução pormenores capelas histórica tipologia exteriores histórico
Cluster 2: àescala mesquitela rapa cadastral sacadas octogonal ratoeira trapalhão chafarizes complementação culinário polidesportivo quadrangular zonamento àfundação judaicas corola subscritos invasoras longitudinais fornotelheiro parcelar alçado topográfica velosa aglomerado poligonal medicinal justificativa esclarecendo limbo plant anexa fotografado topográficos agricultor expropriações longitudinal soleira urbanizável espaços-canais quadrada assinalando arade abóbada operativas telheiro preenchida cúpula volumetria desenhadas arcadas cércea enrolamento toponímia zenith anexas capela-mor cala camionagem delimitação parasita ornamental vigente calvário imune pluviais transversais sacristia fossas urbanizáveis cotas contíguos detenham continha cotada gare àentrada operativa urbanísticos arruamento viária calendarização estacionamentos perímetro ecológica arranjo delimitada exista perpendicular levantamentos telhado indicando claustro alinhamento àcâmara extracto condicionamentos adjacentes subscrito topografia rústico sementeira definindo aglomerados plantação estimativa incorporados cultivo especialidades ratificado específico regulamentares exigente paisagístico edificado abrangendo designadas situa paralelos estatísticos entenda impliquem originária integrante demolição seca construídos licenciadas sanitárias dimensionamento rodar cedência habitacional armazenagem escoamento quiosque rama elaborada telha cortes cresce destinam máximos ecologia variedades filas insectos cave insere viçosa administrativas aterro trate arranjos deter existam mangualde referentes naves transversal irregular escadas muros curvas junção varanda usos futura cadastro aditamento circular comunicar pilares efluentes predominantemente pavimento cota apoiada tronco prestador castanho terço optimização construída detenção vende profundidade lema proteínas moinho garagem estiver
Cluster 3: tubérculos
Cluster 4: ciconia vegetação
Cluster 5: caule folha
Cluster 6: pigmentação
Cluster 7: condicionantes extractos indicação enquadramento existir cópia predial notificação prévia precisa
Cluster 8: rectangular alvenaria
Cluster 9: miróbriga espiga perímetros telhados lusófona primárias arbustos organigrama soeiro linhares grão actualizada mediateca altitude povoado muralha religiosos atalaia muralhas espólio coberta raiz imaginário turística papelaria quintal calçada
Cluster 10: medianamente caules vegetativo susceptibilidade floração polpa batateira parasitas escassa susceptível superficiais abundante amarela lisa porte sementes fresco raízes vegetal fruto
Cluster 11: nemátodos grânulos estufa
Cluster 12: espécimes indígenas espécies
Cluster 13: falésia
Cluster 14: cannabis coca designa estupefacientes
Cluster 15: servidões edificações lotes esgotos parcela drenagem urbanas parcelas anexos recuperação existente alvará edificação restrições ocupação fracção utilidade infra-estruturas eléctrica destinada complementares aplicáveis cobertura destinadas substituição florestais
Cluster 16: míldio semente
Cluster 17: ordenamento plano
Cluster 18: implantação
Cluster 19: escala
Cluster 20: lageosa
Cluster 21: celorico
Cluster 22: descritiva
Cluster 23: síntese
Cluster 24: pisos
Cluster 25: loteamento
Cluster 26: plantas
Cluster 27: perfis
Cluster 28: martelo perna
Cluster 29: pormenor
Cluster 30: pretensão requerente zona
Cluster 31: urbano
Cluster 32: localização
Cluster 33: devidamente
Cluster 34: solos
Cluster 35: poente nascente
Cluster 36: terreno
Cluster 37: urbana
Cluster 38: piso
Cluster 39: operação elementos
Cluster 40: limites
Cluster 41: font
Cluster 42: prédio
Cluster 43: edifícios
Cluster 44: variante
Cluster 45: urbanização
Cluster 46: espaços
Cluster 47: folhas
Cluster 48: urbanos
Cluster 49: chão
Cluster 50: construções
Cluster 51: abastecimento
Cluster 52: solo
Cluster 53: beira
Cluster 54: imóvel
Cluster 55: reserva
Cluster 56: banho
Cluster 57: vias
Cluster 58: rochas
Cluster 59: agrícola
Cluster 60: edifício
Cluster 61: terrenos
Cluster 62: águas
Cluster 63: reabilitação
Cluster 64: conservação
Cluster 65: regulamento
Cluster 66: área
Cluster 67: ribeira
Cluster 68: taxas
Cluster 69: objecto
Cluster 70: eixo
Cluster 71: áreas
Cluster 72: milho
Cluster 73: Áreas
Cluster 74: lote

---

1 clusters com 27 elementos!
1 clusters com 11 elementos!
3 clusters com 3 elementos!
1 clusters com 26 elementos!
1 clusters com 193 elementos!
9 clusters com 2 elementos!
1 clusters com 20 elementos!
55 clusters com 1 elementos!
1 clusters com 4 elementos!
1 clusters com 37 elementos!
1 clusters com 10 elementos!

Anexo C.1
Agrupamentos usado algoritmo CLARA, Informação Mútua e Freq. Min = 2

Cluster 0: sub-paço penisga cruzeiro-dos-pombais ricínio infraes aeisace patótipos ordenamento&raquo patótipo longitudinalp ovulíferas máxima- cilíndrico-cónica pilheiras cmsrp indeiscente erythroxylon ictiotóxico diesel físico-territoriais tuberização folíolo estratgia acúleos georefenciada não-há extracto-resumo alÃ§ados ordenamento-síntese adossando replantando-a reespigamento androsemum rgzedl abadia-mãe todolivo subquadrangular polinervuras rutácea rizoides alkilamidas lapisadas sotano fotogrametrica destine-se envelopante alfizes oligolementos heterospórica hipocolesterolémica referenciando-as castelo-castro desinflamando monogermia electroptécnica viórios vidueiro pedreira- damarana adaptogénica loculicida pre-fabricado glicirrisina liliácea disponibilizalos drosophylum dolaar campanulácea ancÃªde caulescente dicriminação absidadas àescla baptizá-la-ei cimeirinha euforbiacea tetrástila mariamme autonóma pervel âuma adossam aerofotogramético agrotis fasciculados pogressão glandulosos maternalismo crespinho passda cucurbitácea luetzelburgii solânea rea envasada áescala pdmtm infº infestantes-anuais goebelianus &cortes macedinho descontraír absidíola glaucochrous gratissima cartaxaria sementário carrezeda tetrahidrocannabinol não-digital placentiformis patrimón urbanizáveis&raquo estremão &declaração sirgueria imágem eclusão pastado suspeitarmos anserina terphrosia enciclopdia cascone dessecando |levantado pormenor vidre penixobra -what nyack-nyack malvácea montagreste xestion laterítico policrómico thorella buitenland cálcio-fósforo kliex kawipy lilacínea vasaria labiada spotlar ramosa sofá-nido peninérvea pressalit macrotérmicas refernciado fulvilanatus unixessuais primário-vias guribanes doicheman rhine-westphalia oecus monte-de-vénus |anabel chicória-escarola lvcpr carreter scanerizada oceÂnico •clientes lacínias jamacaru parafestuca judicie coll' densiareolatus multicostatus casteleiras leucostele ß-caroteno ex-amada fenÛis intrafascicular peraltados psefito fotogravado lavaÇÃo machadinhos tenreiro-vieira shubo rafflesia mabela arnoldii asturiensis srnorte simularemos coiçe planta-hospedeira arrastÃo transversaiserfpp_ uebelmannia ovado-oblongas si espargo-bravo vialito cmlp quenopodiácea mastodinias cruzarias activadora plantat somniferum rizomatosa lcpv prograding rubiácea reponsÁvel Ésdras adaptogéneo anterídio crucífera
Cluster 1: cortegada alpendrados biselados vernaculares reentrantes biseladas mourela galisteu solarengos quintãs
Cluster 2: prados-
Cluster 3: casa-antiga
Cluster 4: vide-entre-vinhas carvalheda cortiçô maçal carrapichana mesquitela
Cluster 5: síntes
Cluster 6: feira-gastronomia toucinho-céu largo-oliveira
Cluster 7: escravilheira ordasqueira curvel
Cluster 8: oblonga-curta oblonga-alongada amarela-clara semi-precoce branca-amarelada
Cluster 9: sacarrao universit ria
Cluster 10: radicante
Cluster 11: mestre-avis são-lourenço nuno-alvares-pereira asas-abertas limicola
Cluster 12: farol-ponta-altar algar-seco praia-rocha pseudoarmeria
Cluster 13: meio-abertas vermelha-violácea meio-aberta
Cluster 14: matricariae aphidius testaceipes
Cluster 15: isospórica
Cluster 16: topográfico
Cluster 17: gluma
Cluster 18: ecotroca planta-
Cluster 19: cespitosa ráquila
Cluster 20: cilindrico-cónica levogira
Cluster 21: dois-irmãos
Cluster 22: excrecência aristada mucrão lanceolada subiguais
Cluster 23: urbanimétrico microsoft
Cluster 24: bordarias petequino moscatoira
Cluster 25: |situa |proximidade
Cluster 26: bompreÇo galvasud bardella
Cluster 27: glumela
Cluster 28: cariopse
Cluster 29: consolda filodendro oleandro
Cluster 30: sub-erecto
Cluster 31: estolho
Cluster 32: tsukami tokui hachiji
Cluster 33: |apólice |declara|
Cluster 34: galinhatos
Cluster 35: |transversa| |peto |hypeel |elipsóide
Cluster 36: cura exorcismo
Cluster 37: landward lapili
Cluster 38: |rossol |saint-pierre |robin |macero |intensidad| |compriment| |manific |terço |observado
Cluster 39: cimeiras| |carmello |alaranjado |early |cerise |europeel |erlidor |cimeiras |ferline |fruto |amarela
Cluster 40: schilden schopten schopte wisten braken trokken
Cluster 41: praia-do-abano
Cluster 42: plakte vloog verdiende wachtte betaalt dronk
Cluster 43: |gro |plana |floradade |fandango |pedicelar |apla |count |heinz rio| |mech
Cluster 44: frassino alistrong prÉgia aligreen
Cluster 45: híppicos
Cluster 46: vlogen plakten wachtten dronken
Cluster 47: escariosa
Cluster 48: sub-erecta
Cluster 49: |sungold |laranja cor|
Cluster 50: polisacaridos
Cluster 51: |maturação
Cluster 52: mútica
Cluster 53: euforbiácea
Cluster 54: geniculada
Cluster 55: referidosna
Cluster 56: subnace
Cluster 57: |campbell |marmande |monfavet |floração
Cluster 58: aristamento
Cluster 59: panícula
Cluster 60: |planta
Cluster 61: gancheada
Cluster 62: stephanocereus
Cluster 63: pálea
Cluster 64: derris cinerariaefolium
Cluster 65: polymictic polygenic
Cluster 66: venis indusa
Cluster 67: quibla
Cluster 68: nervação
Cluster 69: pré-anotados
Cluster 70: |firme
Cluster 71: àescala
Cluster 72: aerofotogramétrico
Cluster 73: antocianínica
Cluster 74: pilosocereus

---

3 clusters com 6 elementos!
2 clusters com 11 elementos!
9 clusters com 3 elementos!
1 clusters com 9 elementos!
12 clusters com 2 elementos!
1 clusters com 218 elementos!
37 clusters com 1 elementos!
5 clusters com 4 elementos!
2 clusters com 10 elementos!
3 clusters com 5 elementos!

Anexo C.2
Agrupamentos usado algoritmo CLARA, Informação Mútua e Freq. Min = 20000

Cluster 0: ordenamento
Cluster 1: escala
Cluster 2: síntese indicação devidamente operação
Cluster 3: plantas
Cluster 4: pormenor urbano urbanização
Cluster 5: localização terreno urbana edifícios recuperação espaços urbanos abastecimento solo ocupação infra-estruturas
Cluster 6: enquadramento limites restrições utilidade reserva destinada turística destinadas plano áreas território licenciamento planos
Cluster 7: solos
Cluster 8: actualizada
Cluster 9: específico existir referentes complementares aplicáveis regulamento taxas substituição objecto prévia elementos executar aprovação identificação destinados anexo respectivas depósito licença execução
Cluster 10: piso existente imóvel zona vias cobertura edifício reabilitação área lote municipal estacionamento estrada instalações município arquitectura Água construção obras
Cluster 11: font
Cluster 12: prédio
Cluster 13: espécies agrícola águas conservação florestais Áreas caracterização zonas florestal agrícolas
Cluster 14: cópia estiver vigor conforme respectiva legais indicar correspondente pedido sujeitas aplicável respectivo devendo cumprimento disposições referida competentes
Cluster 15: folhas
Cluster 16: chão beira ribeira flores portimão serra torre paredes
Cluster 17: construções
Cluster 18: folha iluminação corte pedra médio sala interior metros exterior solar
Cluster 19: requerente
Cluster 20: histórica eléctrica circular faixa tabela instalação desenho inferior superfície existentes verdes naturais rede expansão água divisão características transformação industriais
Cluster 21: banho
Cluster 22: comunicar precisa decorre publicada marcação obra adequada orgãos animal director espécie dimensão peças constitui termo espectáculos elemento funcional futuro proposta carácter elevado integram normas integrar índice
Cluster 23: terrenos prédios
Cluster 24: notificação
Cluster 25: eixo
Cluster 26: histórico capela aldeia junta casal turístico casas concelho
Cluster 27: flor
Cluster 28: comprimento largura dimensões
Cluster 29: tópicos
Cluster 30: guimarães paços
Cluster 31: ruas museu
Cluster 32: vírus
Cluster 33: pele
Cluster 34: telefónica
Cluster 35: esclarecimentos
Cluster 36: esquerdo
Cluster 37: latina
Cluster 38: conservatória
Cluster 39: inquéritos

---

1 clusters com 17 elementos!
1 clusters com 20 elementos!
2 clusters com 8 elementos!
22 clusters com 1 elementos!
1 clusters com 4 elementos!
1 clusters com 13 elementos!
2 clusters com 19 elementos!
2 clusters com 10 elementos!

2 clusters com 11 elementos!
2 clusters com 3 elementos!
1 clusters com 26 elementos!
3 clusters com 2 elementos!