Clustering de manchetes com K-Means

Rodrigo Carvalho da Silva *
Bruno Orsi Berton †
Anderson Rocha ‡

Abstract

Neste trabalho é proposto um modelo de clustering baseado no algoritmo K-Means. O objetivo é tentar encontrar grupos de manchetes relacionadas, ou seja, temas de manchetes.

1. Introdução

Neste trabalho foi construído um modelo de aprendizado não supervisionado, utilizando o algoritmo K-Means, para solucionar o problema de reconhecimento de grupos de manchetes de jornal. Para implementação do modelo foram utilizadas as seguintes bibliotecas em python: Numpy, SKLearn, Pandas e NLTK. Os dados utilizados foram manchetes de jornal em inglês publicadas entre os anos de 2003 e 2017 na ABC.

As features utilizadas no modelo foram extraidas a partir das manchetes providas utilizando a técinica de frequência de palavras com pesos ajustados a partir da frequência inversa nos documentos (TF-IDF em inglês). A ideia é encontrar termos que aparecem em cada manchete que são importantes para a definicão do assunto ao qual a manchete se refere. Mais detalhes sobre como as features forão extraídas serão apresentados na seção 2. Na seção 3 o modelo é apresentado assim como seu resultado. Na seção 4 uma conclusão sobre os experimentos é apresentada.

2. Extração de Features

O primeiro passo para a extração das features foi identificar o que era importante em cada uma delas para a resolução do problema. Como se trata de um agrupamento baseado na similaridade das manchetes era necessário evidenciar o conteúdo delas.

*Estudante de MC886. **Contato**: rcarvalho.dev@gmail.com **RA**: 147848
†Estudante de MC886. **Contato**: rcarvalho.dev@gmail.com **RA**: 147848
‡Professor de MC886. **Contato**: anderson.rocha@ic.unicamp.br

O primeiro passo foi preprocessar todas as manchetes. Primeiro foram removidos os sinais de pontuação, uma vez que eles não carregam nenhum conteúdo em si. Em seguida foram retirados números. Estes, apesar de carregarem um pouco de informação com eles, são elementos que só fazem sentido em um contexto maior e que poderiam poluir a informação que pretendemos retirar das manchetes. Continuando, foram removidas palavras que são irrelevantes para o conteúdo da manchete, como artigos, preposição e conjunções. Essas palavras carregam pouco valor semantico e podem ser ignoradas sem prejudicar a identificação do tema de uma frase. Por úiltimo foi extraído o radical de cada palavra utilizada, uma vez que variações de uma mesma palavra poderiam prejudicar o agrupamaneto das manchetes. Se três frases diferentes estão falando sobre corrida, por exemplo, mas cada uma delas utiliza uma variação diferente da palavra, gostariamos de conseguir coloca-las dentro do mesmo cluster apesar disso.

Em seguida foi aplicado o TF-IDF para extrair as features. Foi utilizados um dicionário contendo elementos de uma ou duas palavras. Foi decidido por usar palavras ao invés de caracteres pois acreditamos que palavras são elementos mais apropriados para o problema proposto. Grupos de caractéres são muito bons para detectar estilo de escrita e atribuição de autor pois eles ajudam a capturar vícios de escrita como erros de digitação, uso de pontuação e caracteres especias entre outros. Como esse trabalho se trata de clustering baseado no conteúdo das manchetes essas informações não ajudariam a resolver o problema. E por serem manchetes de jornalísticas assume-se que vícios de escrita não estejam presentes no texto. Já utilizando palavras como elementos do nosso dicionário podemos capturar melhor a relação entre elas e o contúdo das manchetes em si.

Aplicando o TF-IDF no corpus obtemos para cada manchete um vetor esparço de features onde cada posição dele possui um valor que indica o peso daquele elemento para a manchete. Cada elemento pode ser formado por uma ou duas palavras adjacentes no texto. Elementos recebem um peso maior se eles aparecem mais vezes em uma manchete, e se aparecem em poucas manchetes. Ou seja, se ele é um termo importante dentro do corpus e destaca uma ou mais

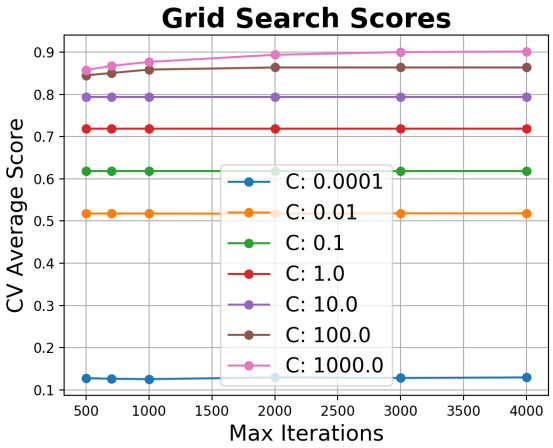

Figura 1. Gráfico da grid search do modelo de Logistic Regression 5-Fold CV

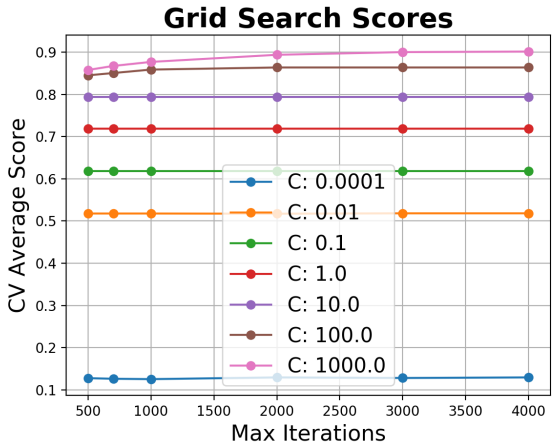

Figura 2. Gráfico da grid search do modelo de Logistic Regression 5-Fold CV

manchetes. Elementos são penalizados se aparecem pouco em uma manchete e se são elementos frequentes dentro do corpus. Ou seja, esse elemento não é importante para a diferenciação de uma ou mais manchetes dentro do corpus.

3. Modelo, Experimentos e Discussão

Após a extração das features o próximo passo foi a aplicação do modelo no nosso dataset. Devido a grande quantidade de dados (Em torno de 1 milhão de manchetes) e a limitação de poder computacional nos computatores utilizados para implementação e testes desse modelo foi utilizada a versão em mini batch do algoritmo K-Means. Onde a cada iteração de treino é selecionado um subset aleatório do dataset e os centroides são atualizados fazendo uma média contínua de cada elemento da mini batch com todos os elementos associados com aos respectivos centroides anteriormente. A versão mini batch de K-Means converge com muito menos custo computacional obtendo resultados geralmente muito próximos aos obtidos pela versão original.

O algoritmo então foi aplicado para valores de K variando entre 2 e 120. Para encontrarmos o valor de K mais apropriado utilizamos duas medidas quantitivas e uma qualitativa. As medidas quantitativas utilizadas foram a função de custo de inercia e o valor shilouette score médio para os clusters. A medida qualitativa foi núvem de palavras.

3.1. Clustering sobre todos os anos

Primeiro foi ralizado o clustering com as manchetes de todos os anos. Na Figura 1 temos a função de custo para os diferentes valores de K.

Na Figura 2 temos o valoe médio do silhouette score médio para os K clusters.

4. Conclusão