Natural Language Processing for Company Business Descriptions: Subcategories and Recommendation

Wei Ding, Wusheng Liu, Kwok Ping Ng
Technical University Munich
wei.ding@tum.de, kwokping.ng@tum.de, wusheng.liu@tum.de

1 Subcategory Analysis

In the last two weeks, we have explored the subcategory of several big categories. We used the unsupervised method K-means. We chose the TF-IDF value as input for K-means instead of Word2Vec, because it is quite straight forward to see the high frequent key words with regard to the TF-IDF value to verify the classification result. Our process flow is:

$$\text{Clean Data} \rightarrow \text{Find Optimal Number of Cluster} \rightarrow \text{Classify Using K-Means} \rightarrow \text{Analyse Results} \quad .$$

We analysed subcategories of 5 groups with relative more data - INFORMATION TECHNOLOGY GROUP, BUSINESS FINANCIAL SERVICES, HEALTHCARE GROUP, CONSUMER SERVICES GROUP, and ENERGY UTILITIES GROUP. In general, we found that the sub-classification for the category 'CONSUMER SERVICES GROUP' is most meaningful. This class has also the most balanced data. For other categories, the results suffer from the imbalance of the data. We analysed the result in Section 1.1 and gave a summary for other categories in Section 1.2.

1.1 Category "CONSUMER SERVICES GROUP"

Table 1 gives an overview of the given information of the Category "CONSUMER SERVICES GROUP". From the table we can observe that this category has four subcategories. Each subcategory has about 200-400 companies. The top keywords for each subcategory are also representative.

Subcategory	# Companies	Top Keywords
Consumer Information Services	407	website,terms,time,content,site
Media and Content	261	education,new,students,learning,content
Travel and Leisure	220	items,service,products,shipping,order
Retailers	201	personal,service,website,services,travel

Table 1: An overview of the Category "CONSUMER SERVICES GROUP" with the given subcategories, number of companies in each subcategory and the top TF-IDF keywords.

Firstly, we need to decide the number of clusters. We put the TF-IDF scores of each website into the k-means classifier. We iterated number of groups from two to ten. Then we got the SSE score from K-means classifier, which calculates the distance between the website and the center of assigned group. So the lower the score is, the better that number of group is. The given number of subcategories is 4. However, from the Figure 1, we can see that K-means suggests that we should choose 8 as the number of clusters for classifying subcategories in "Consumer Services Group".

Figure 2 shows the top TF-IDF keywords after classifying all the companies from "Consumer

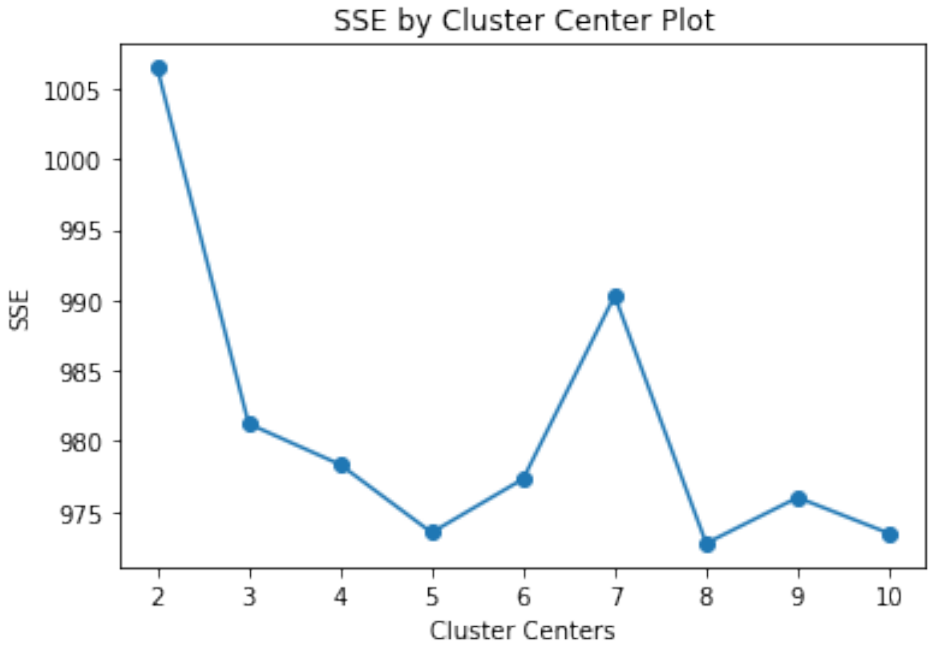

Figure 1: SSE scores for different number of groups for Category "CONSUMER SERVICES GROUP"

Services Group" into 8 subcategories. This classification mainly represented the original 4 given subcategories, while they also explored more details inside the given 4 classes. For instance, under the subcategory "Media and Content", we have now two groups, one is in the education direction, the other is in the content or video direction. Also for the subcategory "Retailers", our classifier finds a group with sweets like candy or chocolate. With the classification, we know more details about the websites in "Consumer Services Group". This classification is quite successful.

Cluster 0 Travel and Leisure
cruise,cruises,river,travelers,trip,tours,tour,trips,stride,travel

Cluster 1 Consumer Information Services
personal,privacy,content,website,terms,service,site,services,use,information

Cluster 2 Consumer Information Services
investment,fleet,instructors,place,funding,business,training,driving,instructor,red

Cluster 3 Retailers
chicago,information,catering,site,beef,llc,hot,theme,dogs,il

Cluster 4 Media and Content
content,powhow,yahoo,team,creators,videos,live,streaming,rdf,video

Cluster 5 Retailers
corp,parkway,com,gourmet,va,la,chocolate,foods,source,candy

Cluster 6 Consumer Information Services
business,like,services,site,service,com,time,use,new,information

Cluster 7 Media and Content
teaching,use,information,university,college,student,school,learning,education,students

Figure 2: Top TF-IDF keywords from classified subcategories

1.2 Other Categories

Other categories are very imbalanced in their subcategories. For example in INFORMATION TECHNOLOGY GROUP, most of websites inclined in Software subcategory. So the clustering results are not as obvious as the "CONSUMER SERVICES GROUP".

Subcategory	#Company
Software	1246
Communications and Networking	222
Electronics and Computer Hardware	218
Semiconductors	89

Table 2: An overview of the Category "INFORMATION TECHNOLOGY GROUP" with the given subcategories, number of companies in each subcategory and the top TF-IDF keywords.

Similarly, the "BUSINESS FINANCIAL SERVICES" does have extraordinary clustering results as well but some noticeable keywords for **insurance industry** from 44 websites. It does not have such subcategory for it. So in this case we can add extra subcategory for insurance industry.

Apart from that, groups sometimes are sharing similar business. For example, the border of the business support service and financial services are quite blur. Their websites have similar keywords then it is hard to cluster them by using TF-IDF.

2 Website Recommendation

We implemented the cosine similarity to each category in the last report. We found that each group has highest similarity value to its own group. So an idea comes up so that we can recommend similar websites to an incoming new website also by using cosine similarity. Our process flow is:

Generate Vectors (tfidf/doc2vec/wordvec) $\rightarrow$Compute Similarity $\rightarrow$Pick 5 Websites with highest Score .

We tested a website "wwww.sbamerica.com" into our recommendation system. This website is promoting healthy food. The contexts of the websites are food and beverage related. Then we used three methods to recommend our existing websites. Here are the results:

Most of the websites recommended by three methods belongs to CONSUMER GOODS GROUP. Among the three methods, word2vec approach gives the best result. We believe the reason is that Word2Vec model uses to generate vector embedding for each word. The embedding is capable of capturing context of a word in a document, as well as semantic and syntactic similarity with other words. As a result, it was able to generate the most similar vectors for words with semantic similar meanings. When calculating cosine similarity of the vectors of the websites top key words, it can give a good representation of the similarity of the business nature of the companies.

In TF-IDF, although most of the websites recommended belong to CONSUMER GOODS GROUP, some websites are not food and beverage related. The first one www.republicind.com is an interior design company, whose area is actually different.

In Word2Vec, all the websites recommended are food and beverage related. For example, www.thanasi.com is a fast-moving consumer-goods company focusing on instant food product.

```
In [181]: #TFIDF Similarity
recommend(Input_Company, top_k, X_tfidf, y_tfidf, Website_tfidf)

Website:  www.republicind.com  Category: CONSUMER GOODS GROUP  Similarity: 0.35
Website:  www.sirkensingtons.com  Category: CONSUMER GOODS GROUP  Similarity: 0.23
Website:  www.wholesomesweeteners.com  Category: CONSUMER GOODS GROUP  Similarity: 0.21
Website:  www.pure360.com  Category: BUSINESS & FINANCIAL SERVICES  Similarity: 0.15
Website:  www.bakewisebrands.com  Category: CONSUMER GOODS GROUP  Similarity: 0.15

In [191]: #Doc2vec Similarity
recommend(Input_Company, top_k, X_doc2vec, y_doc2vec, Website_doc2vec)

Website:  www.lyndale.co.uk  Category: CONSUMER GOODS GROUP  Similarity: 0.91
Website:  www.perosbio.com  Category: HEALTHCARE GROUP  Similarity: 0.91
Website:  www.huskietools.com  Category: INDUSTRIAL GOODS & MATERIALS GROUP  Similarity: 0.91
Website:  www.kerznercareers.com  Category: CONSUMER SERVICES GROUP  Similarity: 0.91
Website:  www.futuresbtc.com  Category: HEALTHCARE GROUP  Similarity: 0.91

In [182]: #Word2vec Similarity
recommend(Input_Company, top_k, X_word2vec, y_word2vec, Website_word2vec)

Website:  www.thanasi.com  Category: CONSUMER SERVICES GROUP  Similarity: 0.92
Website:  www.mainstreetgourmet.com  Category: CONSUMER GOODS GROUP  Similarity: 0.91
Website:  www.bellisiofoods.com  Category: CONSUMER GOODS GROUP  Similarity: 0.91
Website:  www.caesarspasta.com  Category: CONSUMER GOODS GROUP  Similarity: 0.9
Website:  www.rusticcrust.com  Category: CONSUMER SERVICES GROUP  Similarity: 0.9
```

Figure 3: Recommendation for www.sbamerica.com by three methods