Discover and Match Topics

Topic Modeling

14 December 2020

1 Introduction

This exercise is about topic modeling and the confrontation via the Vector Space Model (TF-IDF) and the Latent Dirichlet Allocation (LTA). Via analysis of the given dataset, the query company description has been mapped in order to find the companies best matching with the query company.

2 TF-IDF Topic Modeling

2.1 Cells explanation

In this part of the already completed section of the provided notebook, a TF-IDF based approach has been employed. The snippet described in Figure 1 starts by fitting the companies descriptions to the TF-IDF Vectorizer and, via the fit_transform function, the descriptions are turned into a matrix form.

```
[16]   tfidf = TfidfVectorizer(preprocessor=pre_process).fit_transform(df.description)
```

Figure 1: Cell 10

As an output, a sparse matrix of size 2000×10364 is obtained.

```
[17]   doc_index_to_compare = df.index[df['name'] == "Vahanalytics"].tolist()[0]
       top_k = 5
       cosine_similarities = cosine_similarity(tfidf[doc_index_to_compare:doc_index_to_compare + 1], tfidf).flatten()
```

Figure 2: Cell 11

To explain briefly what the code in Figure 2 does, the dataframe index for the "Vahanalytics" company is obtained and then the variable "top_k" is set to 5. This variable indicates the number of most similar documents to collect. Lastly, the cosine similarities between the "Vahanalytics" company and the others are computed. It is notable the fact that cosine similarities are computed in the matrix space obtained after the execution of the TF-IDF Vectorizer. The matrix space is then flattened into a $1D$ vector via the flatten function.
Figure 3 shows how the similarities are sorted via the argsort function, returning an array of indices, sorted by similarities.

```
[18]   related_docs_indices = cosine_similarities.argsort()[:-top_k - 1:-1]
```

Figure 3: Cell 12

The sorted array is then sliced to obtain the five most similar companies. In Figure 4, by using those indices, the observations carrying those indices are searched in the dataframe.

```
[19]    tfidf_result_df = df[df.index.isin(related_docs_indices)]
```

Figure 4: Cell 13

Since each observation has a unique index, the dataframe created consists of the most similar companies to "Vahanalytics" using the indices found in the previous step. Results are then printed and can be seen in the related provided code.
An identical operation is conducted for the task of finding the most similar companies to the "Much Asphalt" company.

2.2 Much Asphalt Extension

Extending the code cited in the previous subsection to find the most similar companies to "Much Asphalt", description-wise, produces the following interesting results:

- Sunland Asphalt,
- Central-Allied Enterprises,
- FAST FELT,
- Saldus Celinieks

As it can be seen from their descriptions, which are to be read in the provided related code, they are related to the "Much Asphalt" since they all are companies dealing in asphalt, concrete and various other sand and gravel mixes. It's safe to infer, then, that the results do make sense and the Vector Space Model produces solid outcomes.

3 LDA Topic Modeling

In this section, the topic modelling technique based on Latent Dirichlet Allocation has been implemented to match the companies on the basis of company description. Later on, the top 5 closest companies to the query companies "Much Asphalt" and "Vahanalytics" have been found.

3.1 Find the top 5 closest matches (companies) to the companies with names stated

Here presented are the 5 closest companies to Vahanalytics:

- MRM Risk Management,
- 'Koninklijke Mosa',
- 'DynaRoad',
- 'Steven M. Sweat, APC',
- 'Equipment One Stop'

Here are the 5 closest companies to Much Asphalt:

- 'Mover',
- 'Flinders Group',
- 'Premier Logistics Partners',
- 'trans-o-flex Belgium BVBA',
- 'Scotshield Fire Security'

3.2 Which method produces more sensible output? Discuss

During the experimentation phase, it has been observed that the Vector Space Model performed better than the Latent Dirichlet Allocation technique, producing more accurate results. This observation applies for both Much Asphalt and Vahanalytics companies.
TF-IDF Vectorizer is a simple and yet effective model. It better captures semantically closer embeddings, thus models the related topics well within their group.
In comparison, Latent Dirichlet Allocation technique did not perform as well, often producing average results. It is also to be noted that the LDA model requires more hyperparameter tuning than TF-IDF. There is also a limit on the most common tokens to 5000 units. The model that it has been experimented on had a number of topics $K = 10$ and a chunk size equal to 5. One can experiment with these hyperparemeters and might achieve better results since the hyperparameters here used achieved average performance.
Following, a comparison of the retrieved results from both the models. Respectively, first model outcomes belong to TF-IDF and second model outcomes belong to the LDA model. Here are presented the closest companies evaluated by the two methods with respect to the company Vahanalytics:

- 'BISAF: BISAF is a technological company for the construction industry. We specialise in cutting edge solutions that make building easier, safer and environmentally friendly.',
- 'MRM Risk Management: MRM specializes in the evaluation, development, and implementation of wrap-up insurance for large public and private construction programs. They help owners and general contractors achieve simplicity, savings and increased safety for their projects. MRM began as a wrap-up consulting firm but quickly expanded its services to include comprehensive oversight management and then full wrap-up administration in response to their clients' requests and needs'.

Here are the closest companies evaluated by the two methods with respect to the company Much Asphalt.

- 'Sunland Asphalt: Sunland Asphalt, a commercial asphalt paving company in Phoenix, provides commercial asphalt paving service at competitive price',
- 'MOVER: MOVER is a mobile city service for finding and selecting trucks and integrated crossings. The first mobile service that provides services for the organization of complex crossings and cargo transportation in Moscow and the Moscow region.'.

As it can be seen, the different models focused on different common words and, as said before, the TF-IDF model produces better outcomes when compared to the ones produced by the LDA model.