# Ex 06 – Topic Modeling

**Write 1-2 sentences per cell to explain what is done (e.g. explaining the contents of the variables) and why.**

This cell passes the pre-processes description texts to a vectorizer which applies tfidf weighting. The pre-processing steps that are applied can be seen in the cell above in the notebook.

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

I will explain the variables in the cell below:

doc_index_to_compare: get the row index of the row of the company name specified on the right. In this case: Vahanalytics. It first turn the row into a list and then gets the entry at index 0 which is the index of the row in the df.

top_k: just an int defining the number of most similar companies that should be extracted

cosine_similarities: this calculates the cosine similarity for the company description at the given index to all other company descriptions in the list. It does so using the tfidf vectorized representation of the descriptions. As the result is a list of lists, it is flattened

```
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()
```

The cosine similarity array is sorted and the 5 entries with the highest scores are extracted. argsort allows to still keep the original index. This means that these are the indices (of the descriptions) in the df as well.

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

Based on the indices that we got in the cell above, we can now easily extract the row at the resprctive index from the original dataframe.

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

This is a new df, only containing those entries that are in the top 5 most similar entries in the entire dataset.

```
tfidf_result_df
```

**Which are the most similar companies? Do the results make sense?**

The most similar companies for Much Asphalt are the following:

| Company | Description |
|---|---|
| Much Asphalt | Much Asphalt is southern Africa's commercial supplier of an extensive range of hot and cold asphalt products to the road construction economy. Much Asphalt owns and operates 15 static plants in the major centres of South Africa and is the majority shareholder in East Coast Asphalt which operates two more in East London and Mthatha. |
| Sunland Asphalt | Sunland Asphalt, a commercial asphalt paving company in Phoenix, provides commercial asphalt paving service at competitive price. |
| Central-Allied Enterprises | Central States Construction was founded in 1929 by Ernest W. Hallett to produce sand and gravel and construct concrete highways in Minnesota. The business was successful, and in the early 1940s, operations expanded to western Ohio. In the 1940s, the company was heavily involved in the wartime expansion of Wright-Patterson Air Force Base and the post-war construction of the Ohio Turnpike. By the early 1950s, Ohio operations had expanded to include production of sand, gravel, asphalt, and concrete. The Ohio-based portion of the business became known as Allied Enterprises, and it made its permanent presence in Northeastern Ohio by the end of the 50s. Today, Central-Allied Enterprises is one of northeastern Ohio's leading producers of sand, gravel, asphalt, and paved asphalt surfaces. |
| FAST FELT | The patented product FAST FELT®, with its plastic tabs pre-affixed to the asphalt saturated felt (commonly called "tar paper") is the only significant improvement in the recent history of the asphalt saturated felt underlayment products market. |
| Saldus Celinieks | Saldus Celinieks is specialising in road construction, extraction of aggregates and asphalt production. |

Not surprisingly, Much Asphalt itself is among these. This is due to the way the most similar texts are extracted. Although this is not very informative as such, it still tells us that the tfidf vectorizing gives us reasonable results.

## 1. Find the top 5 closest matches (companies) to the companies with names:

Much Asphalt

*NOTE: this table is not the same as in the notebook, I accidentally let it run again, after having interpreted the results.*

| Company | Description |
|---|---|
| Building Structure Institute | Building Structure Institute provides engineering serves including planning, verification, and construction. It offers research planning for seismic isolation, vibration control, seismic isolation technology, experiment verification, and analysis for product development. It's services also includes structural design management. The company was founded in 2010. |
| Miller Supply Inc. | Miller Supply Inc. is a Southern California based packaging distribution company, which has been operating for over thirty years. Miller Supply Inc. specializes in mail center distribtuion and is currently the authorized UPS vendor for Southern Califonria. Miller Supply Inc. also operates the two online ventures Miller Poly Bags & Miller Bubble Mailers, which are product specific distribution websites with a national distribution network. |
| KBB Underground Technologies GmbH | KBB Underground Technologies GmbH, an engineering company, is engaged in planning, building, and operating underground storage facilities and brine/salt extraction facilities. |
| Alliance Wood Group Engineering | Alliance Wood Group Engineering provides engineering, procurement, project management, and construction support services. |
| Burt Hill | Burt Hill Inc. provides architectural and engineering services. Its services include applied research, architecture, engineering, interior design, landscape architecture, planning, sustainable design, and visioning/brand design. The company offers its services for corporate and commercial, destination development, healthcare, higher education, hospitality, K-12, residential, and science and technology projects. |

Vahanalytics

*NOTE: this table is not the same as in the notebook, I accidentally let it run again, after having interpreted the results.*

| Company | Description |
|---|---|
| MechanicNet | MechanicNet Group, Inc. is an innovator of technology and provider of customer retention services specific to the automotive aftermarket. With 14 years of |

| | |
|---|---|
| | leadership in delivering Customer Retention Systems CRM and services to customers including General Motors and Genuine Parts Company. MechanicNet has a proven history of both providing turnkey technologies to Automotive Service Centers as well as managing corporate scale development and deployment projects. In delivering right sized solutions to each customer, they bring to bear a team deep in automotive aftermarket domain expertise, supply chain strategy knowledge, and technology development experience. |
| Concretum Construction Science | The purpose of the company is to provide services (such as consulting, research, system development, training and expertise) in the field of construction materials, as well as the development, production and sale of products for the manufacture and refining of cementitious building materials. The Company may acquire, charge, manage and sell real estate and interests in other domestic and foreign companies. |
| Vahanalytics | Vahanalytics aims to create better drivers and safer roads by using cutting edge big data and machine learning techniques. |
| Data-Maxx Technologies, Inc. | DataMaxx provides a hardware and software solution for tracking time and attendance, equipment, inventory, assets and production. Specializes in field collection for mobile workforce data and provides various construction time clock methods for collecting information. Various device hardware and software systems available for collection of data, including cell phones, PDAs, bar code readers, RFID, biometric fingerprint and hand punch devices, PC time clock and daily report software available. Integrates with most job cost accounting packages. |
| Procsim Consulting | Procsim is at the heart of the EPFL Innovation Park, an ecosystem where mingle research and entrepreneurship, emerging and developing high-tech start-ups. Procsim evolves in a dynamic and extremely stimulating environment, ideal for sharing ideas and knowledge development. |

**2. Which method produces more sensible output? Discuss. (Identify manually based on the description of companies)**

I think it is hard to judge which method worked better. The lda produced very different results if I let it run several times. Sometimes they were really good, sometimes they were rather bad. The results for Vahanalytics based on the tfidf weighting did not at all include some companies connected to roads or drivers. They were all mainly based on technology. Especially the last result (index 1982) is rather off. It mentions ‘machine learning’ in the description and I think this is why it is among the top 5 results. Apart from that, the two companies do not have a lot in common based on what I can judge from the descriptions. The results produced by the lda

are only slightly better. There is a least one results (MechnicNet) that has something to do with cars. However, when looking at the tokenized version of the description, it makes sense that the results are rather bad. For Vahanalytics it contains only 4 words. I tried it once with including most of the words instead of just 5000 to see if there is a difference, but there is not a large difference. For Much Asphalt, the results are slightly better, I think at least for the tfidf. The result descriptions for the lda do have something to do with construction. Still, I know that in the dataset there are companies that are more similar to Much Asphalt (see the new table for Much Asphalt in the notebook, the company Saldus Celinieks is very similar I think). The results produced by the tfidf are quite good. They have something to do with asphalt and road construction and are very similar to the comparison description.

For some runs, the lda produced very good results so I think that if the model was trained on more data it would give very sensible output. Setting only 10 topics might have been too little. This is something that I could further experiment with to improve the results. However, it is good to know that also simple models like tfidf weighting can produce good results.

**Resources used other than the links on the Assignment:**

https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/