Lab Report of Exercise 6

Approaches

Part I Tf-Idf Method

1. Preprocess and vectorize the company descriptions

Apply pre_process function to the description data when using Tf-idf to vectorize them, preprocessing includes lower the case, remove website, email, punctuations and stopwords, stem the words.

2. Calculate the cosine similarity between query company and all companies

Use the Tf-idf vectors of company descriptions to calculate cosine similarity between companies, and sort the similarities and get the top 5 similar companies' indices, then return their information.

Part II LDA Method

1. Preprocess and tokenize the company descriptions

Use the same pre_process function to preprocess the description data, then use simple_preprocess from the gensim module to tokenize each description.

2. Find common tokens and remove uncommon tokens

Define functions to find the top 5000 common tokens and remove the tokens that are not the common tokens from each tokenized description.

3. Create corpus and id2word dictionary for topic modeling

Apply corpora.Dictionary to tokenized description to get the id2word dictionary and get the corpus for LDA modeling.

4. Train models with different num_topics

Try different number of topics and get the best LDA model by measuring the coherence score.

5. Get the LDA results and convert them to 2D arrays

Put the corpus into the selected model and get the results, then convert them to 2D arrays.

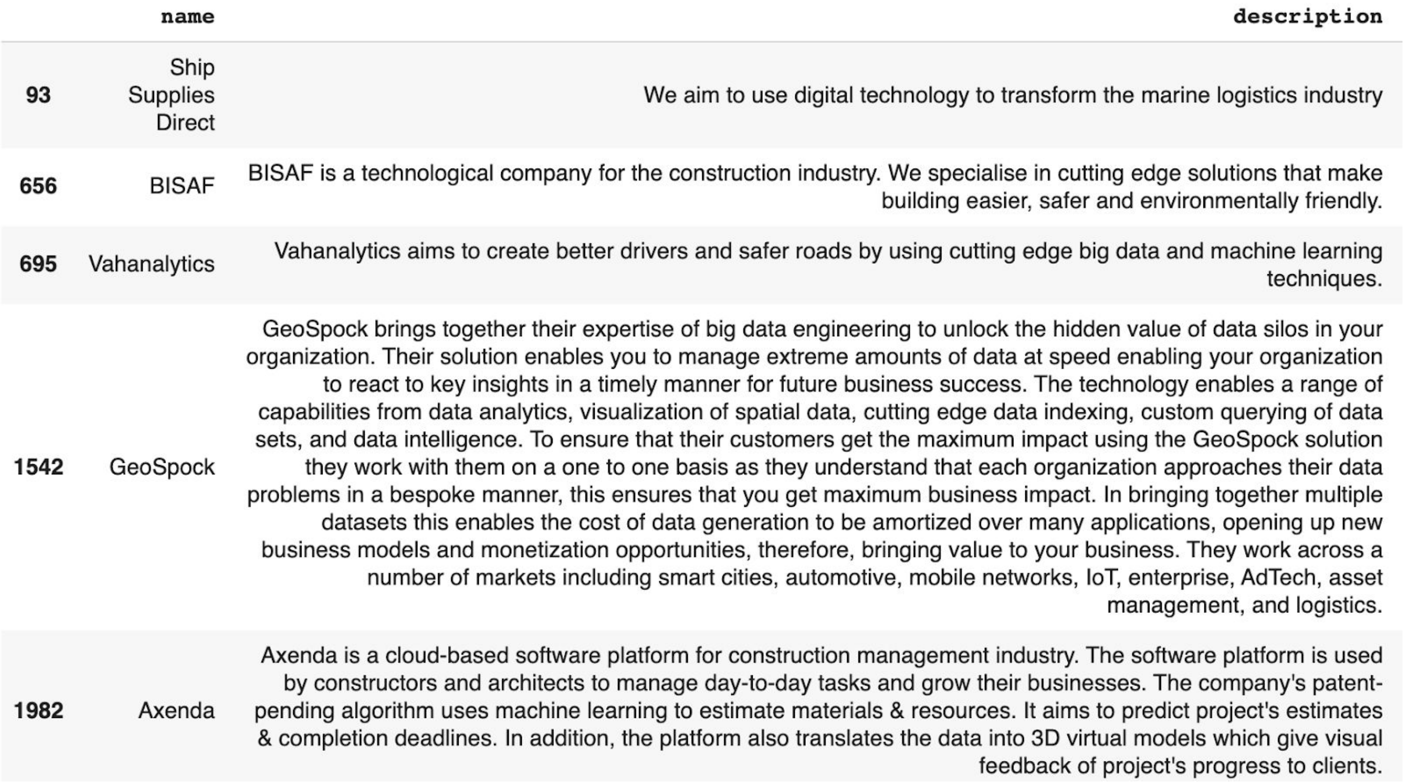
6. Get the top 5 similarly companies to each query company

Get the index of the query company, then we can get the LDA result of this company;
Calculate the jensen shannon similarity between the query company and all companies in the data.
Sort the similarities and get the top 5 closest ones to query company.

Result Comparison

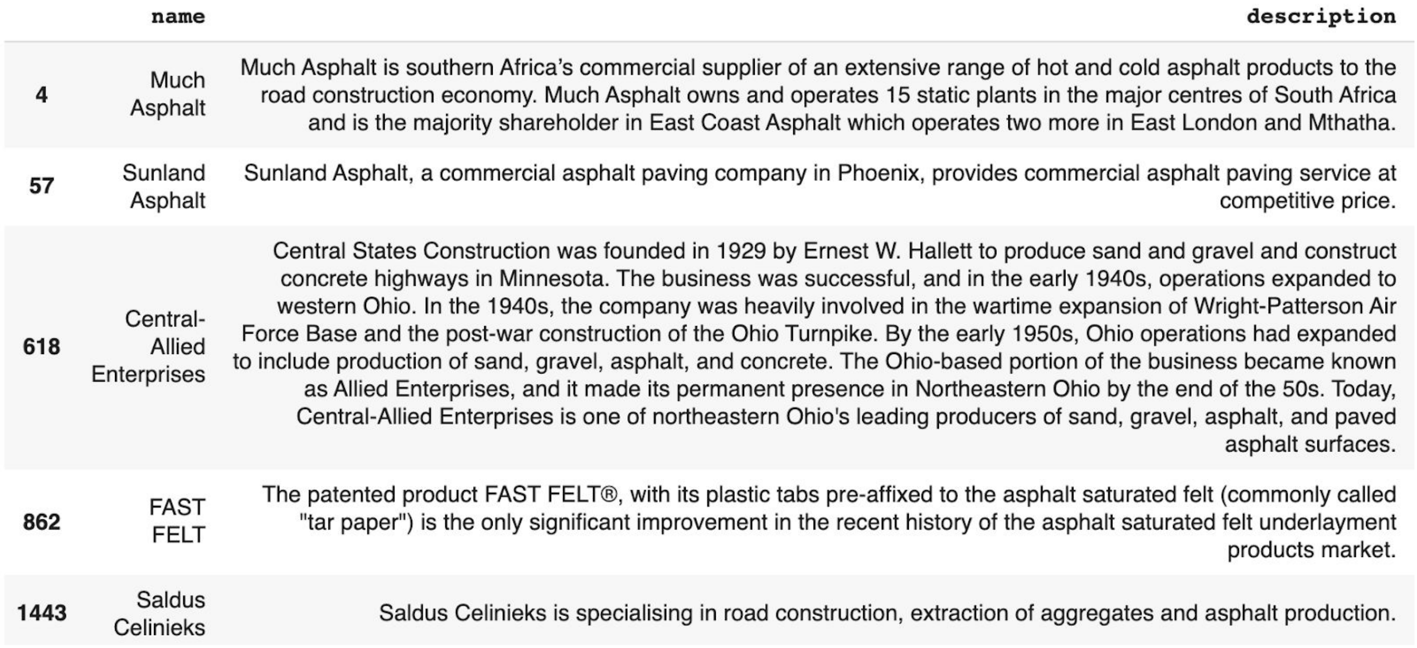
I listed the results of the two methods below, for further discussion, please see the section of Discussion.

Top 5 similar companies to 'Vahanalytics' (including 'Vahanalytics') found by Tf-Idf method



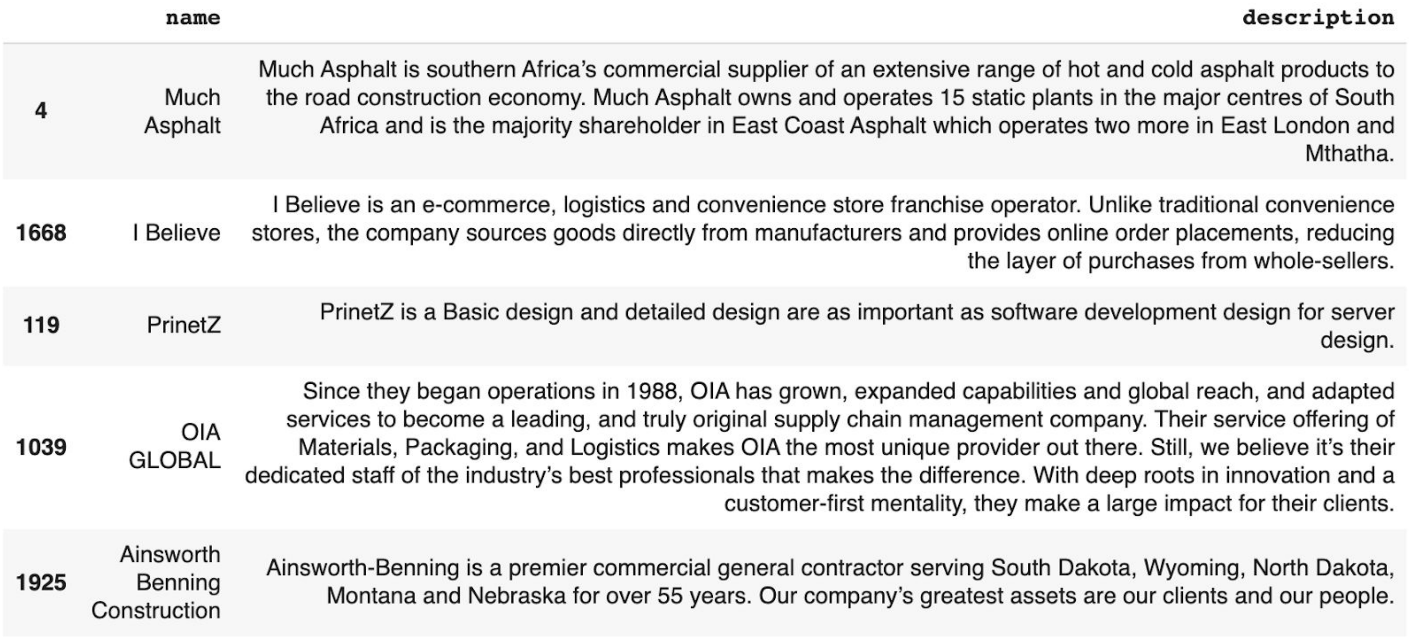
	name	description
93	Ship Supplies Direct	We aim to use digital technology to transform the marine logistics industry
656	BISAF	BISAF is a technological company for the construction industry. We specialise in cutting edge solutions that make building easier, safer and environmentally friendly.
695	Vahanalytics	Vahanalytics aims to create better drivers and safer roads by using cutting edge big data and machine learning techniques.
1542	GeoSpock	GeoSpock brings together their expertise of big data engineering to unlock the hidden value of data silos in your organization. Their solution enables you to manage extreme amounts of data at speed enabling your organization to react to key insights in a timely manner for future business success. The technology enables a range of capabilities from data analytics, visualization of spatial data, cutting edge data indexing, custom querying of data sets, and data intelligence. To ensure that their customers get the maximum impact using the GeoSpock solution they work with them on a one to one basis as they understand that each organization approaches their data problems in a bespoke manner, this ensures that you get maximum business impact. In bringing together multiple datasets this enables the cost of data generation to be amortized over many applications, opening up new business models and monetization opportunities, therefore, bringing value to your business. They work across a number of markets including smart cities, automotive, mobile networks, IoT, enterprise, AdTech, asset management, and logistics.
1982	Axenda	Axenda is a cloud-based software platform for construction management industry. The software platform is used by constructors and architects to manage day-to-day tasks and grow their businesses. The company's patent-pending algorithm uses machine learning to estimate materials & resources. It aims to predict project's estimates & completion deadlines. In addition, the platform also translates the data into 3D virtual models which give visual feedback of project's progress to clients.

Top 5 similar companies to ‘Much Asphalt’ (including ‘Much Asphalt’) found by Tf-Idf method



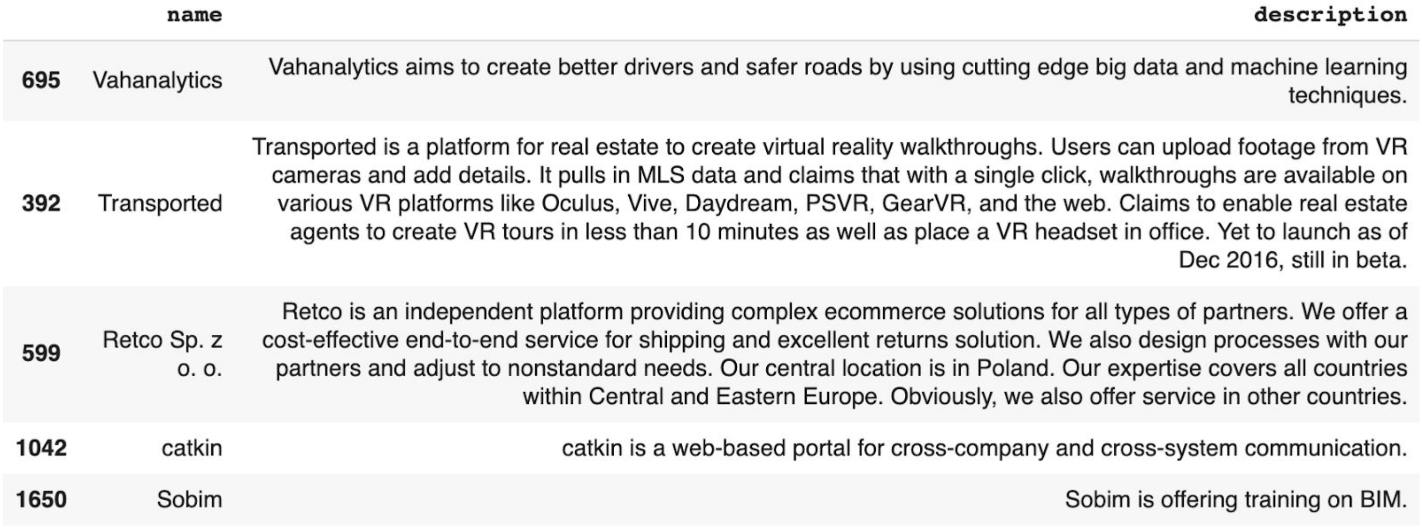
	name	description
4	Much Asphalt	Much Asphalt is southern Africa’s commercial supplier of an extensive range of hot and cold asphalt products to the road construction economy. Much Asphalt owns and operates 15 static plants in the major centres of South Africa and is the majority shareholder in East Coast Asphalt which operates two more in East London and Mthatha.
57	Sunland Asphalt	Sunland Asphalt, a commercial asphalt paving company in Phoenix, provides commercial asphalt paving service at competitive price.
618	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.
862	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.
1443	Saldus Celinieks	Saldus Celinieks is specialising in road construction, extraction of aggregates and asphalt production.

Top 5 similar companies to ‘Much Asphalt’ (including ‘Much Asphalt’’) found by LDA method



	name	description
4	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.
1668	I Believe	I Believe is an e-commerce, logistics and convenience store franchise operator. Unlike traditional convenience stores, the company sources goods directly from manufacturers and provides online order placements, reducing the layer of purchases from whole-sellers.
119	PrinetZ	PrinetZ is a Basic design and detailed design are as important as software development design for server design.
1039	OIA GLOBAL	Since they began operations in 1988, OIA has grown, expanded capabilities and global reach, and adapted services to become a leading, and truly original supply chain management company. Their service offering of Materials, Packaging, and Logistics makes OIA the most unique provider out there. Still, we believe it’s their dedicated staff of the industry’s best professionals that makes the difference. With deep roots in innovation and a customer-first mentality, they make a large impact for their clients.
1925	Ainsworth Benning Construction	Ainsworth-Benning is a premier commercial general contractor serving South Dakota, Wyoming, North Dakota, Montana and Nebraska for over 55 years. Our company’s greatest assets are our clients and our people.

Top 5 similar companies to ‘Much Asphalt’ (including ‘Much Asphalt’’) found by LDA method



	name	description
695	Vahanalytics	Vahanalytics aims to create better drivers and safer roads by using cutting edge big data and machine learning techniques.
392	Transported	Transported is a platform for real estate to create virtual reality walkthroughs. Users can upload footage from VR cameras and add details. It pulls in MLS data and claims that with a single click, walkthroughs are available on various VR platforms like Oculus, Vive, Daydream, PSVR, GearVR, and the web. Claims to enable real estate agents to create VR tours in less than 10 minutes as well as place a VR headset in office. Yet to launch as of Dec 2016, still in beta.
599	Retco Sp. z o. o.	Retco is an independent platform providing complex ecommerce solutions for all types of partners. We offer a cost-effective end-to-end service for shipping and excellent returns solution. We also design processes with our partners and adjust to nonstandard needs. Our central location is in Poland. Our expertise covers all countries within Central and Eastern Europe. Obviously, we also offer service in other countries.
1042	catkin	catkin is a web-based portal for cross-company and cross-system communication.
1650	Sobim	Sobim is offering training on BIM.

Discussion

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

Please see the python notebook for explanations of the cells.

Table 1. Variables and Their Content

Variables	**Content**
tfidf	Preprocess the description data, and vectorize the company descriptions using the method Tf-Idf, the content of this variable is the vectorized descriptions.
doc_index_to_compare	The index of chosen company
cosine_similarities	A list of cosine similarities between the chosen company and all the companies in the dataset
related_docs_indices	The indices of the top 5 closest companies of the chosen company
tfidf_result_df	Based on the companies’ indices, return the information of the 5 closest companies in the dataset

(2) Extend the code in this section to get the top 5 most similar companies to the company "Much Asphalt". Which are the most similar companies? Do the results make sense?

Please see Top 5 similar companies to 'Much Asphalt' (including 'Much Asphalt') found by Tf-Idf method.

The most similar companies to 'Much Asphalt' are 'Much Asphalt', 'Sunland Asphalt', 'Central-Allied Enterprises', 'FAST FELT' and 'Saldus Celinieks'. By analyzing the descriptions, we can see that the result is quite good. 'Much Asphalt' is a supplier of Asphalt, 'Central-Allied Enterprises' and 'Saldus Celinieks' also Asphalt producers; 'Sunland Asphalt' is a asphalt paving company, although it's not a asphalt producer, it has something to do with asphalt as wee. 'FAST FELT' is a patented product, this one is a bit far from an asphalt producer.

Table 2. Top Similar Companies to 'Much Asphalt' Found by Tf-Idf Method

Name	**Type of the Company**
Much Asphalt	Commercial **supplier of Asphalt** in South Africa
Sunland Asphalt	Commercial Asphalt paving company in Phoenix
Central-Allied Enterprises	**Producer of** sand, gravel, concrete, **asphalt** and paved asphalt surfaces
FAST FELT	A patented product, its plastic tabs pre-affixed to the asphalt saturated felt
Saldus Celinieks	road construction, extraction of aggregates and **asphalt production**

(3) Which method produces more sensible output?

Now we compare the results of the two methods. From Table 3 we can see, the result of the Tf-Idf method is much better than the result of LDA. Table 4 is the comparison of the top 5 similar companies to 'Vahanalytics', both methods didn't return good results, but the Tf-Idf method is a bit better than LDA.

LDA performs worse than Tf-Idf is probably due to (1) LDA is better for large corpus; (2) In this case, the descriptions are all about companies, words are quite specific in meaning. When using the TF-IDF method, the model focuses on the keywords, by comparing the keywords, it's more accurate to match similar companies.

Table 3. Comparison of the Top 5 Similar Companies to 'Must Asphalt'

Tf-Idf		**LDA**	
Much Asphalt	Commercial **supplier of Asphalt** in South Africa	Much Asphalt	Commercial **supplier of Asphalt** in South Africa
Sunland Asphalt	Commercial Asphalt paving company in Phoenix	I Believe	E-commerce, logistics and convenience store franchise operator
Central-Allied Enterprises	**Producer of** sand, gravel, concrete, **asphalt** and paved	PrinetZ	Basic design and detailed design

	asphalt surfaces		
FAST FELT	A patented product, its plastic tabs pre-affixed to the asphalt saturated felt	OIA GLOBAL	Supply chain management company
Saldus Celinieks	Road construction, extraction of aggregates and **asphalt production**	Ainsworth Benning Construction	Commercial general contractor

Table 4. Comparison of the Top 5 Similar Companies to ‘Vahanalytics’

Tf-Idf		**LDA**	
Ship Supplies Direct	Use **digital technology to transform the marine logistics** industry	Vahanalytics	Create better drivers and safer roads by using cutting edge **big data and machine learning techniques**
BISAF	Technological company for the construction industry, make building easier, safer and environmentally friendly	Transported	Platform for real estate to create virtual reality walkthroughs
Vahanalytics	Create better drivers and safer roads by using cutting edge **big data and machine learning techniques**	Retco Sp. z o. o	Platform providing complex ecommerce solutions
GeoSpock	**Service of big data analytics**	catkin	Web-based portal for cross-company and cross-system communication
Axenda	Cloud-based software platform for construction management industry	Sobim	Offer training on BIM