UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Master thesis (30 ECTS)

Entity linking via topic models in Apache Spark

Olha Kaminska

Supervisors:
Pelle Jakovits, PhD
Peep Küngas, PhD

June 5, 2019

Introduction

- **Entity linking** is a field of natural language processing that aims to define the real meaning of a word in a particular text.
- The same **term** can have different meanings in different contexts, which demonstrates the importance of the field.
- Entity linking is actively applied to real-world **business problems**.

Problem statement

- The **purpose of the entity linking** in the current work is to detect the similar products among companies assortment.
- This thesis was created based on the needs of the **company Register OÜ**.
- The task of the current work was an improvement of entity linking quality presented in **a master's thesis**, created by Madis-Karli Koppel for Register OÜ in 2018 year.

Related work

Large Scale Feature Extraction from Linked Web Data
Madis-Karli Koppel (University of Tartu, 2018)

Koppel created the pipeline, where **the input** was linked data from web crawling, and **the output** was features for credit-scoring model. The **steps** of pipeline:

1. Open linked data were extracted via web crawling.
2. Ambiguity was eliminated using named entity recognition.
3. A linked graph was created, where the links were based on stock keeping units (SKUs) of products.
4. Features for credit-scoring model were extracted from the graph.

The outcome was the successful implementation of pipeline, but with **an issue in the entity linking** part.

The main goal

- Koppel used products with **SKUs**, but only 1.9% such products were connected to some companies.
- **The goal** of this thesis is an investigation whether it is possible to use topic modelling to improve the number of connected companies.
- In other words, is it possible **to improve** entity linking using topic-modeling methods.

Pipeline

1. Select products with **textual descriptions**, detect language, choose the major ones.
2. Perform **text cleaning**, select from product descriptions only nouns using POS-tagging technique.
3. Separate product descriptions based on topics, obtained with four different **topic modeling techniques**. Products considered to be similar are clustered together.
4. **Build graphs** for each language and for each topic modeling method separately. Products in the same cluster are linked among each other.
5. Compare obtained graphs to define **the best method** of topic modeling.
6. **Compare** the best graph with Koppel's to detect did it improve the previous result.

Source data

- Source data about **450,096 products** were extracted via crawling Estonian open web in August 2017 and provided by Register OÜ.
- For every product, among all 59 characteristics, the **textual** ones ("Name", "Description" and "Product id") were joined and used as input for topic modeling techniques.
- The number of instances with **English** joined text is 187,036 products, with **Estonian** - 134,332 products.

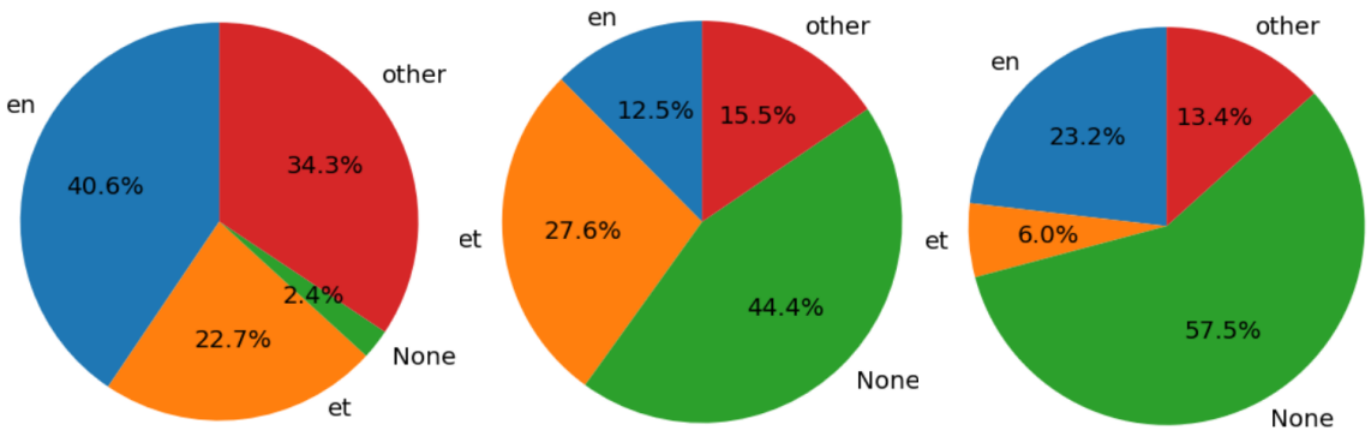


Figure: Language distribution for textual columns.

Topic modeling methods

1. The **latent Dirichlet allocation** (LDA), in two forms: based on a corpus embedded with a bag-of-words model, and with a TF-IDF.
2. The **DBSCAN algorithm**, where product descriptions firstly are transformed into vectors using the Doc2Vec approach.
3. Keyword extraction with a **KMeans clustering algorithm**, where for each product, the words in the product description are presented as vectors using the Word2Vec.
4. Keyword extraction with **TF-IDF scores**, where for each product description, the words with the largest TF-IDF scores are extracted as keywords.

Graphs building and validation

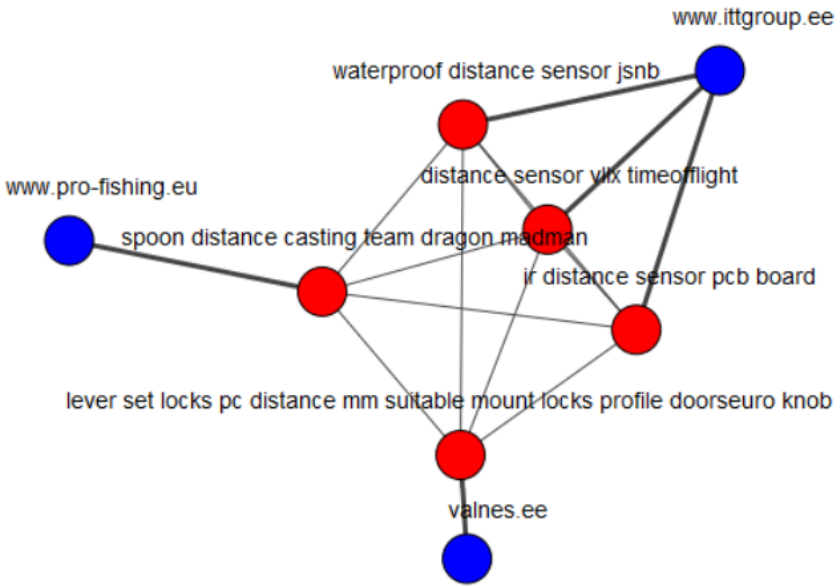
- A relationship graph linking companies through products was built separately for each language and for each topic modeling method. As output, **10 graphs** were obtained.
- **Apache Spark** was chosen as an appropriate environment because it allows users to work with a large number of data quickly.
- At present, no one universal approach for graph comparison exists. In this thesis, **several criteria for graph evaluation** have been proposed.

Graphs comparison I
Statistical characteristics

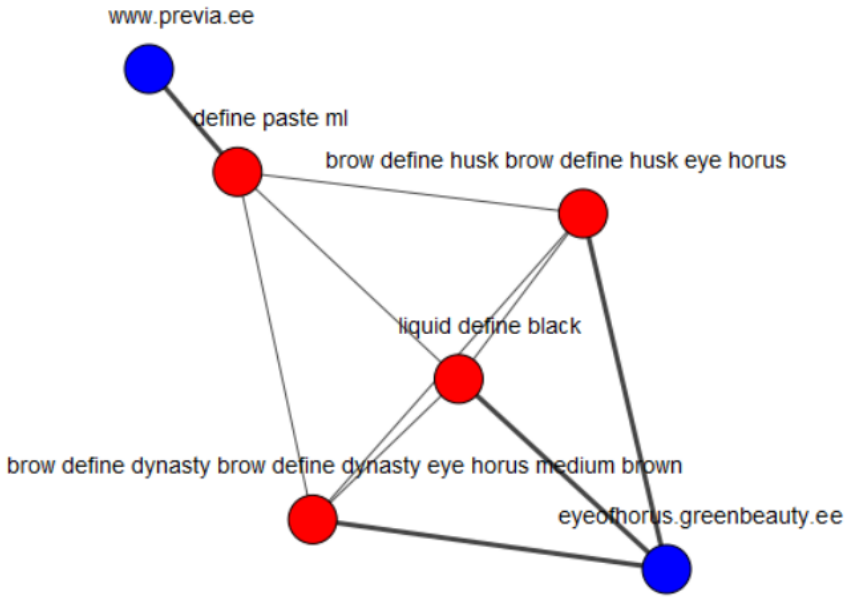
Characteristics	Methods				
	LDA (BOW)	LDA (TF-IDF)	TF-IDF	Kmeans	DBSCAN
Amount of clusters	54	54	9,155	30,536	626
% of links	11.2%	10.6%	3.0%	4.5%	0.1%
Min prod amount	1,225	1,545	1	1	10
Mean prod amount	3,452	3,452	58	19	96
Max prod amount	33,450	30,922	14,718	13,932	1,593

Figure: Comparison of graphs based on different clusters sets for English.

Graphs comparison II
Subgraphs samples



(a) With TF-IDF.



(b) With KMeans.

Figure: Example of graph elements for English data based on keywords extraction.

Graphs comparison III
Comparison of products between two similar companies

Methods	Characteristics				
	Linked nailin.ee	Linked biotrend.ee	Linked amoung all	Samples linked	Samples non-linked
LDA (BOW)	76.3%	98.8%	86.2%	"acrylic color powder" "pigment rosa"	"metallic pastel pink" "compact makeup"
LDA (TF-IDF)	74.0%	96.6%	83.9%	"professional acrylic paint" "rainbow top coat"	"gel brush flat dark" "oil lavender skin"
Kmeans	72.1%	85.4%	77.9%	"sunny orange" "vitamin bomb"	"smaller bullionsbullions" "glitter multicolor"
TF-IDF	53.8%	61.3%	57.3%	"hand repair cream" combination skin cleansing gel"	"beautiful confetti flakes" "sample body"
DBSCAN	21.8%	50.6%	30.4%	"hand antiseptic" "nail art detail"	"berry juice" "combination skin night cream"

Figure: Measurement of connected products from two similar companies.

Graphs comparison IV
Comparison of products between two different companies

Methods	Characteristics				
	Linked smarta.ee	Linked all4pet.ee	Linked amoung all	Samples linked	Samples non-linked
LDA (BOW)	17.5%	13.6%	15.4%	"otterbox commuter series case"	"fashion case iphone aurora red" "royal canin kitten"
LDA (TF-IDF)	15.8%	94.5%	57.7%	"acana dog pacifica"	"magnet wallet iphone plus black" "orijen dog puppy"
Kmeans	11.3%	72.5%	43.9%	"royal canin jelly"	"startpakett basic" "orijen dog senior"
TF-IDF	2.9%	46.1%	25.9%	"hillubs feline adult ocean fish"	"magnet wallet iphone" "brit care cat missy"
DBSCAN	16.6%	6.3%	9.5%	"royal canin mini exigent"	"start kit monitor" "large breed lamb rice"

Figure: Measurement of connected products from two different companies.

Summary

- Based on the graphs comparisons results, the **LDA approaches** were the best.
- **Keyword extraction** based on KMeans clustering and TF-IDF scores did not generate very good results.
- **DBSCAN** failed to present high-quality results because it labeled more than half of the data as noise.
- **The best topic modeling approach** was chosen the BOW-based LDA approach, it created 11.2% of all possible links among all products.

Comparison with Koppel's results

- As a basis for the linking, **Koppel used SKUs**, but only 1.9% of the products had them.
- In the current thesis, **textual information about the products** was chosen as the source data for the linking. Near 71.4% of the products had either English or Estonian description.
- The obtained results **significantly increased** the number of links from Koppel paper.
- Important to note, that the graph from the previous work connected only products that **are exactly the same**. The graph in the current work connects products that **belong to the same topic**.

Conclusion

- The work was performed with **the purpose** of improving the entity linking in the pipeline presented by Koppel thesis using topic modeling techniques. **BOW-based LDA approach** was chosen the best topic modeling approach.
- The main **contribution** of this thesis is the methodology for applying topic modeling to improve entity linking. It was confirmed that the methodology **increased** the number of connected products (71.4%) in comparison to the previous work (1.9%).
- The current work **provided** a graph that can be used to generate more features. Whether the additional links are usable for the original use case of the company, needs more investigation and evaluation by the domain experts.