KU LEUVEN

Topic modeling applied on innovation studies of Flemish companies

Master Thesis in Data Science
Statistics Flanders Data Science Event
May 24th, 2022

Victor Vanhullebusch & Annelien Crijns

Promotor: Prof. Bart Baesens
Supervisor: Dr. Michael Reusens
Daily Supervisor: Ms. Manon Reusens

Agenda

1. Topic & Purpose
2. Related Research
3. Research Questions
4. Research Methodology & Results
    1. Data
    2. Web Scraping & Storage in Database
    3. Preprocessing & Language Detection
    4. Topic Modeling: Top2Vec
    5. Classification Model: Lbl2Vec
5. Conclusion

Agenda

1. **Topic & Purpose**
2. Related Research
3. Research Questions
4. Research Methodology & Results
    1. Data
    2. Web Scraping & Storage in Database
    3. Preprocessing & Language Detection
    4. Topic Modeling: Top2Vec
    5. Classification Model: Lbl2Vec
5. Conclusion

Topic & Purpose

Our thesis contributes to better statistics about innovation in Flemish companies through the use of web scraping and text mining

Task:
- Describing & mapping innovation in companies

Traditional approach:
- Community Innovaton Survey (CIS)
- Every 2 years, all EU Member States

Drawbacks:
- Only companies with 10+ employees
- Lack of motivation to respond (correctly)
- High costs
- Slow process

New approach:
- Web-scraping of business websites
- Text mining

Advantages:
- Small companies included
- No response burden
- Cheaper
- Faster

Result:
- Official statistics about innovation on Flemish, Belgian, European scale

Agenda

1. Topic & Purpose
2. **Related Research**
3. Research Questions
4. Research Methodology & Results
    1. Data
    2. Web Scraping & Storage in Database
    3. Preprocessing & Language Detection
    4. Topic Modeling: Top2Vec
    5. Classification Model: Lbl2Vec
5. Conclusion

Related Research

Previous research shows relevance of web scraping, but focuses on binary classification

Web Scraping

- **Downloading, parsing, and organizing data** from the **web** in an **automated** manner (vanden Broucke & Baesens, 2018).
- Research shows that web scraped data can be **reliable for reporting**

Innovation Studies

- Previous research **about innovation detection via web-scraping of company websites** in the UK, the Netherlands, Germany, Flanders
- Often focus on reproducing CIS results: **classification of companies into innovative or non-innovative**

Our Research

- **Mapping innovation** through **web scraping and text mining**

Difference with previous publications:

- Innovative combination of two recently released, complementary models: **Top2Vec and LbI2Vec**
- **Detection of topics** in Flemish, innovative companies
- **Classification** of companies into the **discovered topics**

Agenda

1. Topic & Purpose
2. Related Research
3. **Research Questions**
4. Research Methodology & Results
    1. Data
    2. Web Scraping & Storage in Database
    3. Preprocessing & Language Detection
    4. Topic Modeling: Top2Vec
    5. Classification Model: Lbl2Vec
5. Conclusion

Research Questions

Our main goals is to find out how to leverage topic modeling in the detection of innovation in Flemish companies

"To what extent can **topic modeling** enhance **detection of innovation in Flemish companies**?"

"What are the main **topics or sectors** represented in Flemish, **innovative** companies"

"How can companies be **classified** into the **discovered topics or sectors**, based on their scraped business websites?"

"What **words** on company websites are most related to **innovation**?"

Agenda

1. Topic & Purpose
2. Related Research
3. Research Questions
4. **Research Methodology & Results**
    1. **Data**
    2. Web Scraping & Storage in Database
    3. Preprocessing & Language Detection
    4. Topic Modeling: Top2Vec
    5. Classification Model: Lbl2Vec
5. Conclusion

Research Methodology: Data

CIS 2019 data from 3,179 Flemish companies is used

Data source:
- CIS 2019 results
- 3,179 Flemish companies participated
- Company name, BTW number, address, email, website, and inno5

Inno5
- Binary variable: innovative or non-innovative
- Based on CIS questions about 5 types of innovation:

Product innovation

Business process innovation

Abandoned innovation

On-going innovation activities

In-house or external R&D activities

Attention points:
- Missing values
- Class imbalance: 75% innovative
- Only a sample, not all Flemish companies are included

Agenda

1. Topic & Purpose
2. Related Research
3. Research Questions
4. **Research Methodology & Results**
    1. Data
    2. **Web Scraping & Storage in Database**
    3. Preprocessing & Language Detection
    4. Topic Modeling: Top2Vec
    5. Classification Model: Lbl2Vec
5. Conclusion

Research Methodology: Web Scraping & Storage in Database

URLs are scraped via Requests and BS4 libraries, and stored in ArangoDB

Web Scraping

Task:
- Gather web texts from given URLs in an automated manner

Approach:
- Check if URLs are valid via “get” requests
- Scrape available and valid URLs
  - Requests and Beautiful Soup 4 (BS4) libraries

Result:
- Gathered all web texts without manually opening URLs and copying data

Storage in Database

Task:
- Storing dataset and scraped web texts

Database: ArangoDB (community edition 3.8.4)
- Open-source database system
- Access and modify via Python
- Database visible via web interface

Result:
- Stored dataset in key/value pairs
  - Primary key: BTW number
  - Values: firm name, address, zip code, email, website, inno5
- Stored scraped texts in key/value pairs
  - Primary key: unique number
  - Values: firm name, website, scraped text

Agenda

1. Topic & Purpose
2. Related Research
3. Research Questions
4. **Research Methodology & Results**
    1. Data
    2. Web Scraping & Storage in Database
    3. **Preprocessing & Language Detection**
    4. Topic Modeling: Top2Vec
    5. Classification Model: Lbl2Vec
5. Conclusion

Research Methodology: Preprocessing & Language Detection

Standard preprocessing with customized stopword removal was applied, and most web texts are in Dutch or English

Task:
- Prepare text for text mining techniques
- Detect languages of scraped web texts

Approach:
- Removed texts without relevant content
  - Texts with 20 or less words
  - Domain names for sale
- Language detection through langdetect library in Python
- Cleaning
  - Tokenization, Lowercase, Punctuation removal, Removal of numbers, Removal of text less than three characters
- Stopwords removal
  - Standard stop words (Dutch) →"NLTK" library
  - "Company website stopwords" (related to cookies, opening hours, language choice) → created list based on frequently observed words

Result:
- 95%+ Dutch or English → texts in other languages removed
- After scraping and preprocessing 2,096 firms (65.9%) remain

Agenda

1. Topic & Purpose
2. Related Research
3. Research Questions
4. **Research Methodology & Results**
    1. Data
    2. Web Scraping & Storage in Database
    3. Preprocessing & Language Detection
    4. **Topic Modeling: Top2Vec**
    5. Classification Model: Lbl2Vec
5. Conclusion

Topic Modeling through Top2Vec (1)

Nine clear topics were detected in the innovative companies' scraped web texts

Task:
- Identify the **main topics** occurring in the **innovative** companies' web texts
- Identify the **words** occurring on company websites that are most related to **innovation**

Approach: Top2Vec model
- Released in 2020 by Angelov, D.
- Semantic multidimensional space is created
  - Word and document vectors
  - Distance represents semantic similarity
- 20 iterations
  - Stochastic model
  - 10 on Dutch Innovative dataset, 10 on English Innovative dataset

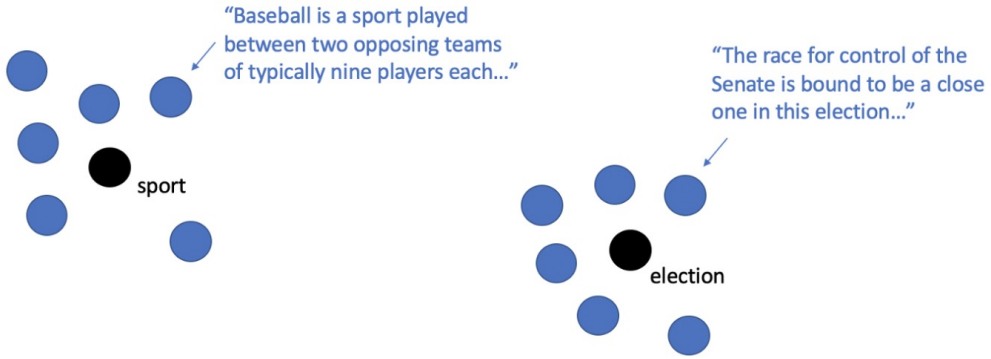
Result: 9 economic sectors / themes were discovered
- Output word clouds, representing clear topics
- Topics were linked to economic sectors described by MVO Flanders

Example of an output word cloud representing the Transport & Logistics sector

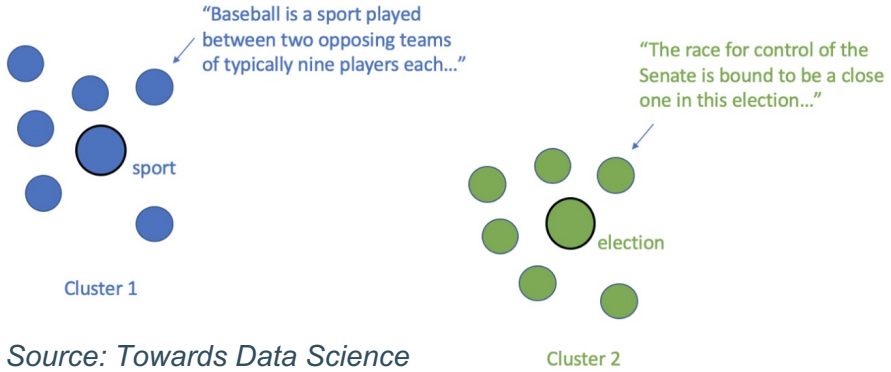
Topic Modeling through Top2Vec (2)

Simple visualisation of the Top2Vec algorithm

1. Word and document vectors

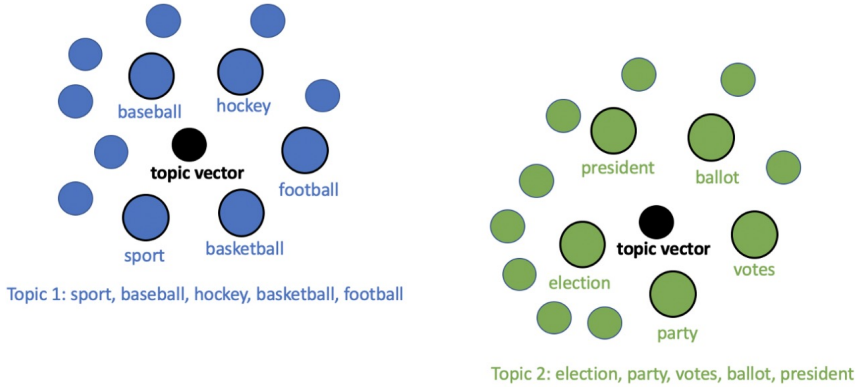


2. Clustering word and document vectors



Source: Towards Data Science

3. Topic clusters and keywords



Topic Modeling through Top2Vec (3)

Eight topics were linked to economic sectors

Dutch & English innovative dataset

Transport and logistics

Food

Engineering

Wood and furniture

IT and consultancy

Dutch innovative dataset

Banking and insurance

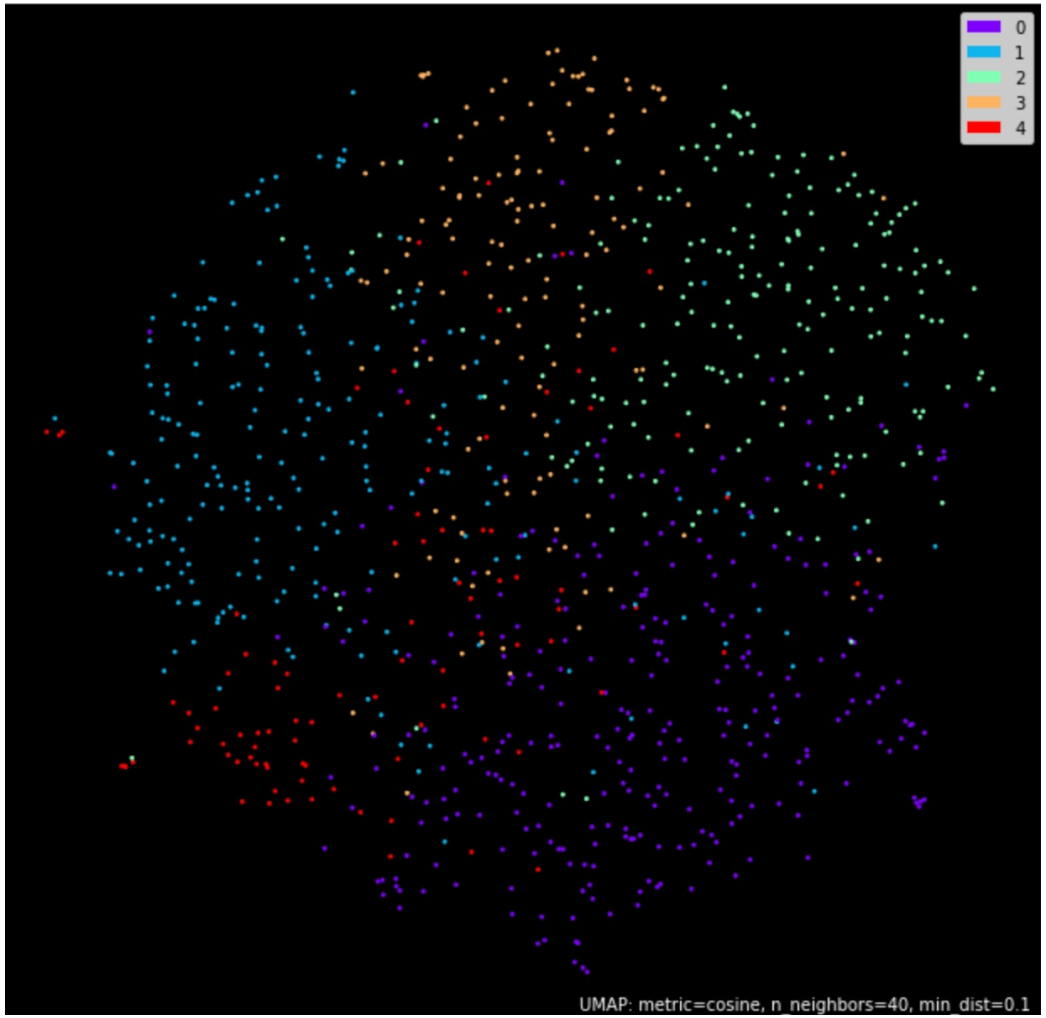
English innovative dataset

Textile and clothing

Health

Topic Modeling through Top2Vec (4)

UMAP displays the points from a higher dimensional space to a lower one (5 to 2)

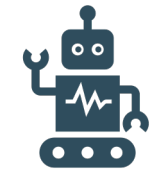


Topic Modeling through Top2Vec (5)

Words closely related to innovation include technology, sustainability, and future

sustainability 0.4419727051419598
sustainable 0.40329130469687896
technology 0.40295140119324097
impact 0.37657928107521776
strategy 0.3738219256432806
investors 0.3704678110520671 6
future 0.36131394954353346
human 0.355270982978517
global 0.3548269487969177
careers 0.3538855759386136
stories 0.3488221489271728
collaboration 0.3339649081068234
compliance 0.3210649473202607
locations 0.3146686418589482
history 0.3094542988004745
governance 0.3090988143510286
report 0.3078667763233083
reports 0.3069678878959824
vision 0.3051306587786596
industry 0.3007615103790799

energy 0.300717838285083 25
solutions 0.2985318136536784
approach 0.2897486761213238
progress 0.28730947450483263
insights 0.2860226472349644
research 0.2823912395038255
strong 0.28161665268363956
challenges 0.2803518868462526
climate 0.2779307772178241
electronics 0.2757308972919674
life 0.2751585294854305
corporate governance
0.2737768794713682
opportunities 0.27359574952774424
commitment 0.273379604437859
values 0.2710405541277462
overview 0.267919123998673 74
supply 0.267673842999659
businesses 0.266675053351483
newsroom 0.26655423531260963
world 0.26655120449927217



Topic Modeling through Top2Vec (6)

Top2Vec can display the company most related to innovation

Most innovative company according to the model:

international optical optoelectronic technology international products solutions international products solutions diagnose treat medical technology assure industrial quality excellence industrial metrology spectroscopy researching micro nanostructures microscopy enabling chip manufacturers semiconductor manufacturing technology see better vision care photograph film photography cinematrography internationally leading technology enterprise operating fields optics optoelectronics solutions contribute world optics development technological progress passion excellence create value customers inspire world see new ways glance develop produce oem solutions digital solutions software development create virtual worlds simulation solutions planetariums observe experience hunting nature observation innovation technology international innovation technology megatrends trends perspectives innovation research research award colloquium symposium promoting innovation deutscher zukunftspreis new business digital partners ventures innovation hub kit international company glance management committees carl foundation annual report years history glance carl ernst abbe technological milestones archives optical museum museum optics year moon landing anniversary sustainability sustainability sustainability strategy products value chain integrity compliance employees engagement environment system certifications latest news international latest news events events calendar forum day light newsroom latest news press releases press photos press kits press contacts press conferences stories discover world optics careers international careers career get overview discover areas research development sales service digital innovations project management production manufacturing

Agenda

1. Topic & Purpose
2. Related Research
3. Research Questions
4. **Research Methodology & Results**
    1. Data
    2. Web Scraping & Storage in Database
    3. Preprocessing & Language Detection
    4. Topic Modeling: Top2Vec
    5. **Classification Model: Lbl2Vec**
5. Conclusion

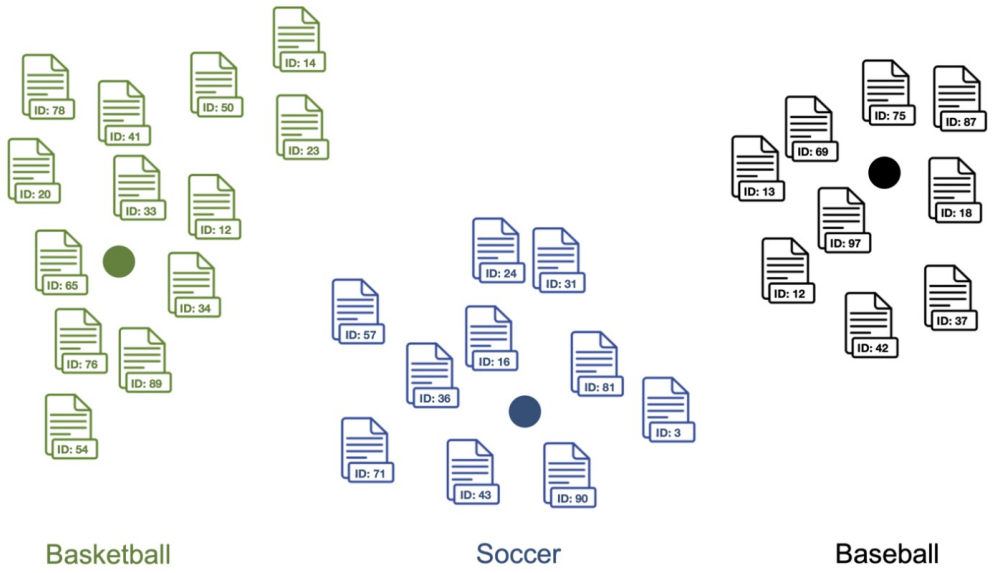
Classification of firms into topics through Lbl2Vec

The Lbl2Vec model was successful with an overall classification accuracy of 75%

Task:
- **Classify** the companies **into the topics** which were discovered through Top2Vec

Approach: Lbl2Vec model
- Released in 2021 by Schopf T., Braun, D. & Matthes, F.
- Unsupervised
  - No labeled data needed for training → time and cost saving
- Semantic multidimensional space is created, just like Top2Vec
  - Word, document, and label vectors
  - Distance represents semantic similarity
- Input
  - Topics and accompanying topic words → output from Top2Vec is used as input
- Best performance when high intratopic and low intertopic similarity

Result:
- Tested on ~300 manually labeled firms and 6 classes
- Overall accuracy of 75%, and F1-scores ranging from 0.60 to 0.88



Classification results for all documents in the dataset. Points represent label vectors of a classification category. Document colors represent their predicted classification category. Image by author.

Source: Towards Data Science

Agenda

1. Topic & Purpose
2. Related Research
3. Research Questions
4. Research Methodology & Results
    1. Data
    2. Web Scraping & Storage in Database
    3. Preprocessing & Language Detection
    4. Topic Modeling: Top2Vec
    5. Classification Model: Lbl2Vec
5. **Conclusion**

Conclusion

The combination of Top2Vec and Lbl2Vec models can be successful in innovation studies

Top2Vec & Lbl2Vec as Complements

Top2Vec
- **Detection of topics** in a large unstructured text corpus
- Identified **economic sectors** in which Flemish innovative firms are active

Lbl2Vec
- **Unsupervised classification model,**
- Successfully classified ~300 firms with an overall accuracy of 75%
- **Used Top2Vec output** topics and words as input

→ **Innovative combination of two recently released models**

Further Research

- Further explore the possibilities of **topic detection** and classification in **innovation studies**

- Conduct on **larger scale** including more companies
  - More topics and related keywords in Top2Vec
  - More exhaustive inputs for Lbl2Vec,

  → More complete classification including **classes representing all economic sectors**

Value & Implications

- Value in further research on developing **more complete statistics**

- **Example: company database** with the possibility to filter firms based on economic sector, geographic location etc.
  - Finding firms for partnerships
  - Analysing competition
  - …

Thank You!

Q&A