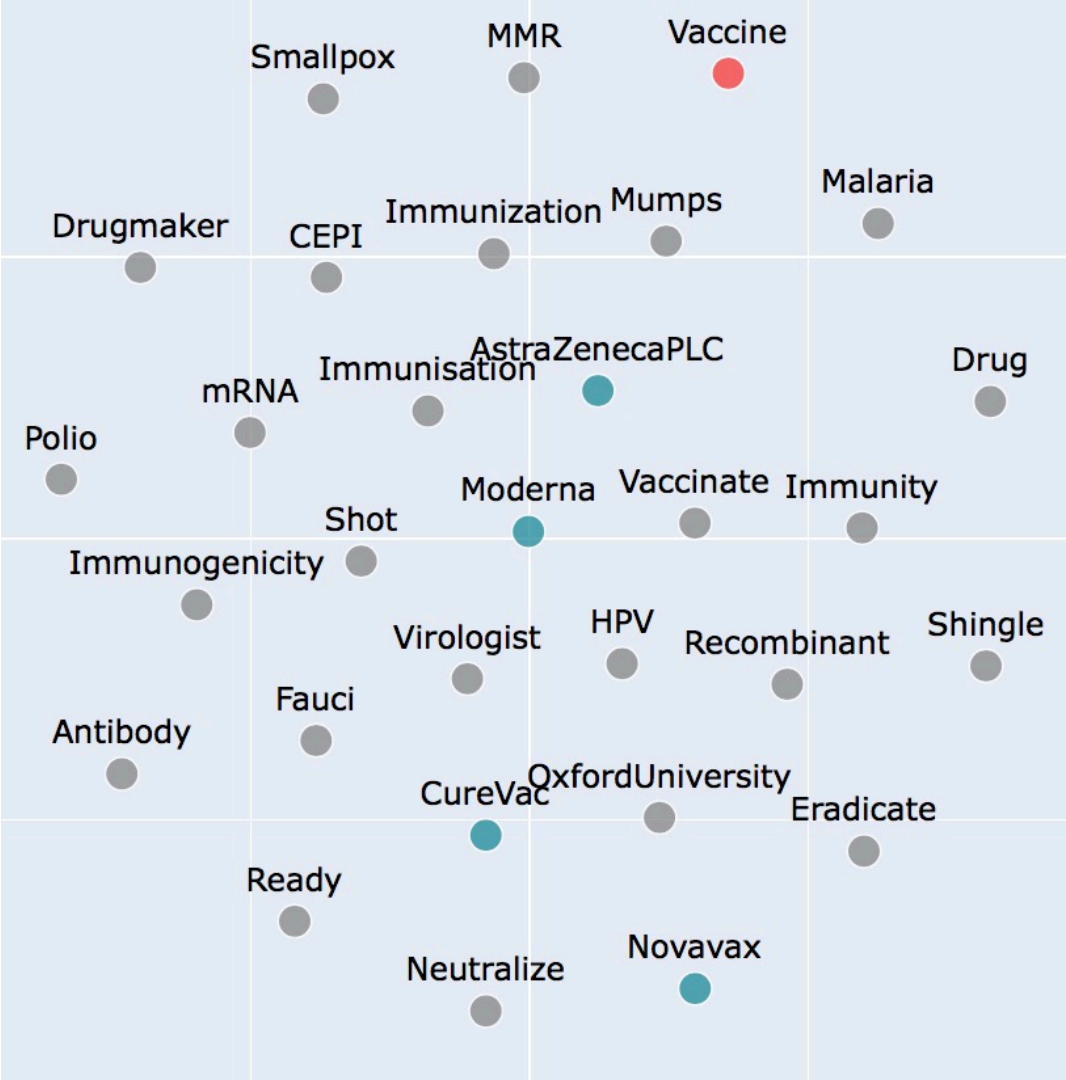
Extracting insights from the word vector space

Janna Lipenkova
Janna.Lipenkova@anacode.de

ARIC Brown Bag Session, July 28, 2020



OUTLINE

anacode

1. Challenge: B2B research

2. Basics of word vectors

3. Demo: analyzing Covid-19 vaccine companies

anacode

B2B research is done on a daily basis

Identify and analyze companies for:

- Customer acquisition
- Supply
- Cooperation
- Competitive intelligence
- Investment
- …

anacode

The traditional approach

Google & Co.

covid vaccine companies

All | News | Images | Shopping | Videos | More | Settings | Tools

About 337.000.000 results (0,63 seconds)

Ad · www.who.int/myth_busters/vaccines

Vaccines against pneumonia do - not protect against COVID-19

No. **Vaccines** against pneumonia, such as pneumococcal **vaccine** and Haemophilus influenza. type B (Hib) **vaccine**, do not protect against the coronavirus. Learn more. Recommendation from WHO. Advice for the public. Official WHO website.

www.gtai.de › press › germany-coronavirus-vaccine-research-231420

German companies at the heart of coronavirus vaccine efforts ...

Jul 1, 2020 - There is cautious optimism that biotechnology and pharmaceutic know-how from Germany can help produce a **COVID**-19 **vaccine** in the near ...

www.forbes.com › sites › moneyshow › 2020/06/16 › 9-pharmaceuti...

9 Pharmaceutical Companies Racing For A COVID-19 Vaccine

Jun 16, 2020 - Globally, there are more than 100 **vaccines** under development — with nine of them in human clinical trials already, asserts equity analyst Sel ...

www.biopharmadive.com › news › coronavirus-vaccine-pipeline-types

The coronavirus vaccine frontrunners are advancing quickly ...

Jun 9, 2020 - Use the dropdown to highlight events in each **company's** timeline. Solid dots indicate events which have occurred, while striped bars indicate ...

www.marketwatch.com › story › these-nine-companies-are-working-on-...

These 23 companies are working on coronavirus treatments ...

May 6, 2020 - GlaxoSmithKline. Type: **Vaccine**, treatment. Name: AS03 adjuvant system for **vaccines**. Background: GlaxoSmithKline US:GSK is another leading ...

www.statista.com › ... › Pharmaceutical Products & Market

• COVID-19 drugs vaccines in development top companies ...

As of July 22, 2020, there were 487 drugs and **vaccines** in development targeting the

Structured databases

Current search:

- Freetext: covid-19 vaccine

Results 1 - 20 out of 46 displayed. Order by: Relevance Sort direction: ▲

1 2 3 »

BioNTech SE (100%)
Mainz, Germany

BIONTECH

Main sector:
- Biotechnology - Therapeutics and Diagnostics

Subsector:
- Anti-infectives
- Antibodies
- Cell therapy
- Gene therapy
- Immunotherapy
- Nucleic acid drugs
- Proteins
- Small molecules
- Stem cells
- Vaccines

Indications:
- Infectious and parasitic diseases / infectiology / parasitology
- Neoplasms / cancer / oncology
- Respiratory / pulmonology
- Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

Employees:
- Worldwide: 1100

anacode

NLP-based approach

- Large dataset of online texts (news, social media, patents, ...)
- Updated in real-time
- Processed with Natural Language Processing
- Focus on word vectors (high-dimensional representations of concepts, incl. companies)

OUTLINE

1. Challenge

2. Basics of word vectors

3. Demo: analyzing Covid-19 vaccine companies

anacode

„Know a word by the company it keeps“ (Firth 1957)

- Distributional similarity: similar words appear in similar contexts.

  The customer finally signed the cotnratc.

  Synonyms for contract: *agreement, deal, arrangement…*

- Different kinds of semantic similarity:

  Topical: *dog, barked, leash*

  Categorical: *Poodle, Pitbull, Rottweiler*

  Syntactic: *walking, running, sprinting*

anacode

Word representations

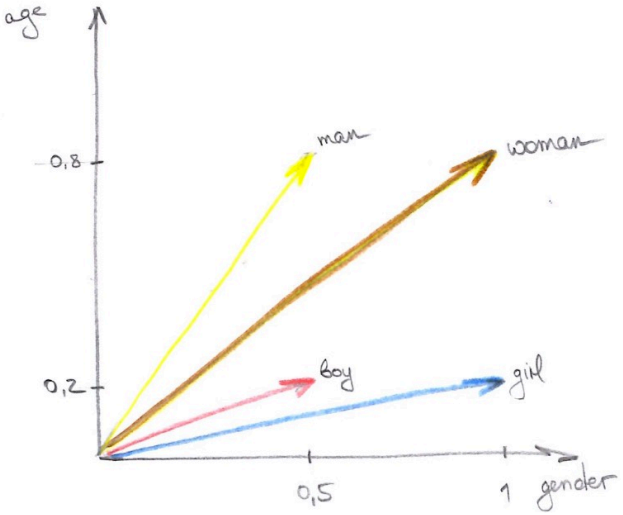
1. One-hot representation (old-school):

Boy	1	0	0	0
Girl	0	1	0	0
Man	0	0	1	0
Woman	0	0	0	1

- > 100k dimensions
- Doesn't capture similarities between words

2. Distributed representation (word vectors):

	(age)	(gender)
	0.2	0.5
	0.2	1
	0.8	0.5
	0.8	1

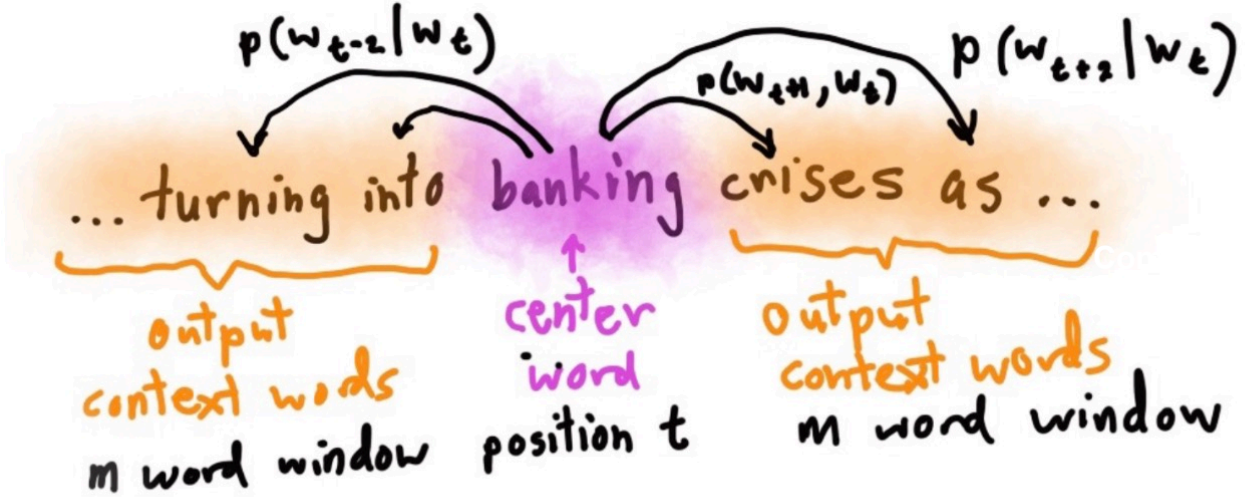


- 100-1000 dimensions
- Captures feature-based similarities

anacode

Training word vectors

- Unsupervised learning on text data
- Given a center word w_t: predict its context words in window of specific size:



OUTLINE

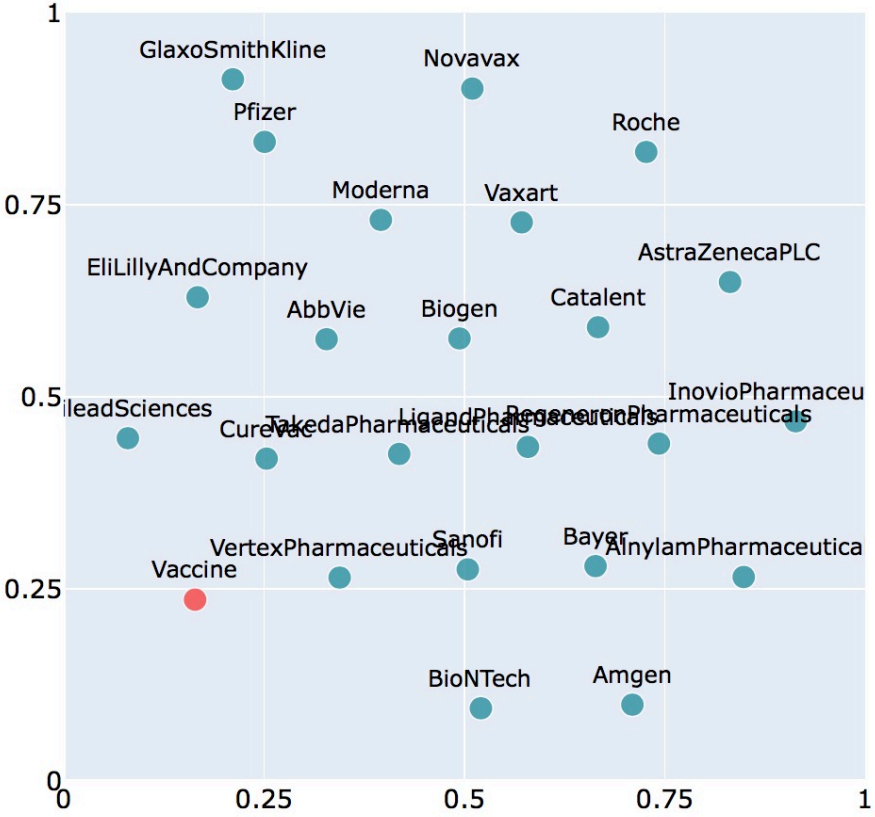
1. Challenge

2. Basics of word vectors

3. Demo: analyzing Covid-19 vaccine companies

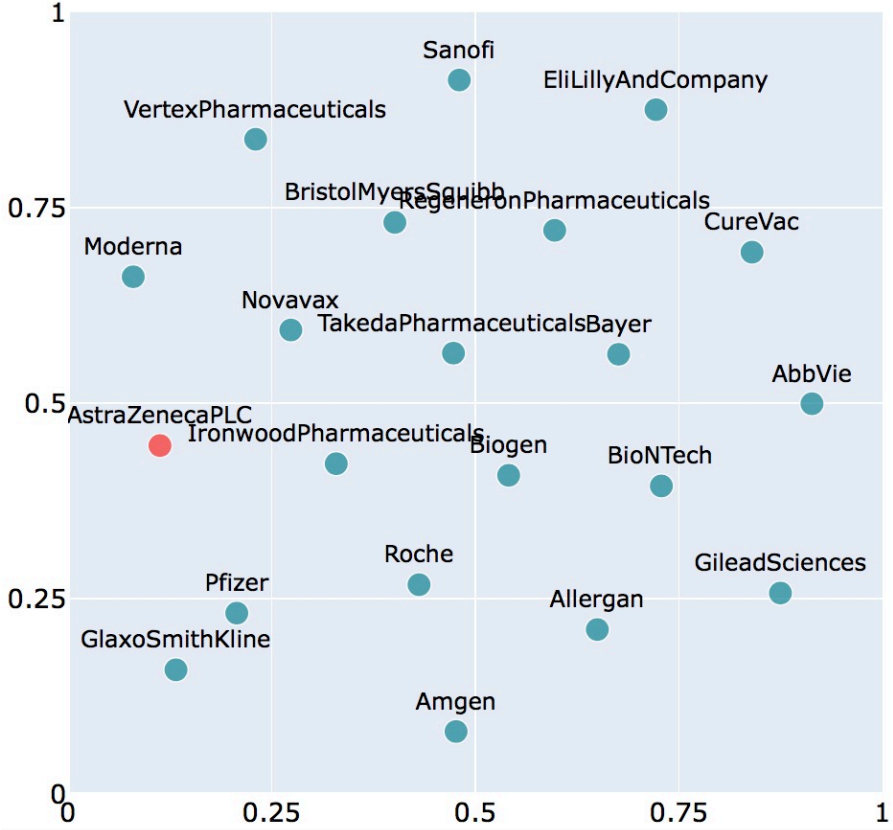
anacode

Vaccine-related companies



anacode

AstraZeneca PLC and peers

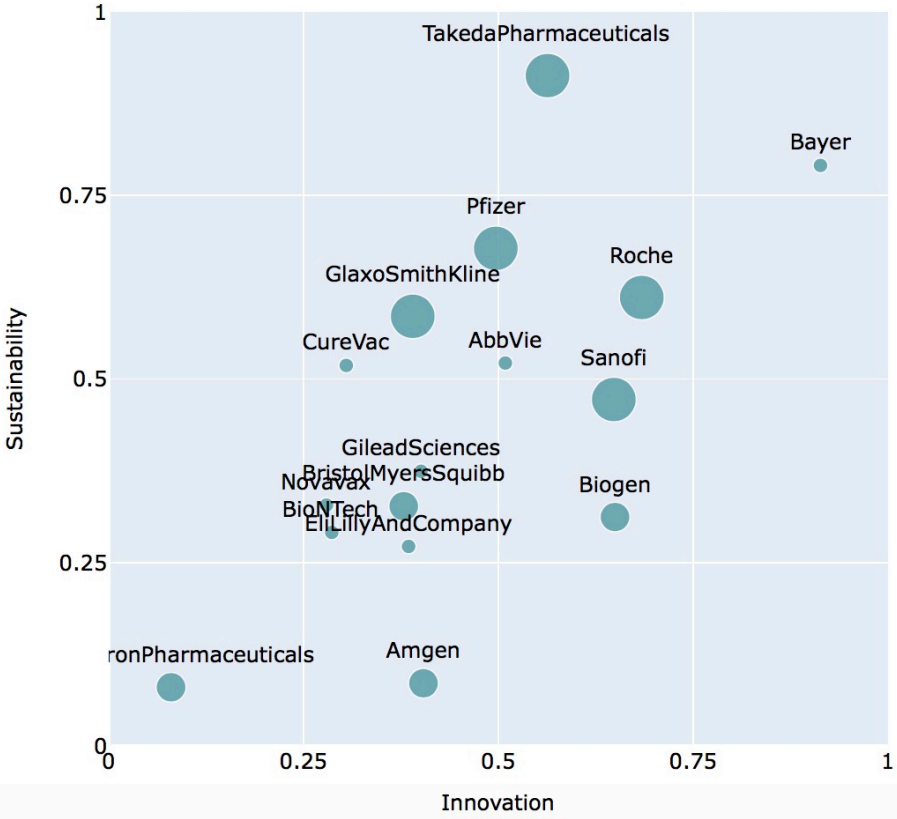


anacode

Giving meaning to x and y

X-Axis Innovation

Y-Axis Sustainability



anacode

Take-aways

- Word vectors allow to detect similarities and relations between concepts.
- The detection is fully unsupervised – thus, highly efficient.
- Word vectors capture the complete semantic value of a text corpus.
- Myriad of use cases, incl. search, recommendation, classification, construction of knowledge bases...

anacode

Curious about NLP?

- Read our Review of NLP research in past 20 years (shorturl.at/efjoG)
- Check out our Covid-19 Public Media Dataset and train your own word vectors!
- Presubscribe to my blog on NLP and Advanced Analytics: drop an e-mail with subject line „subscribe“ to nlpblog@anacode.de.

CONTACT

Anacode GmbH
Kurfürstendamm 76
10709 Berlin

info@anacode.de

www.anacode.de

Dr. Janna Lipenkova
+49 152 098 17 228