The ABC of Computational Text Analysis

#8 Ethics and the Evolution of NLP

Faculty of Humanities and Social Sciences
University of Lucerne

22 April 2022

Faculty of Humanities and Social Sciences

Recap last Lecture

- an abundance data sources

  JSTOR, Nexis, few datasets

- creating your own dataset

  converting any data to `.txt`

Outline

- cover last lecture's batch processing 
- ethics is your responsibility 🙈🙉🙊
- understand the development of modern NLP 🚀

  … or how to put words into computers

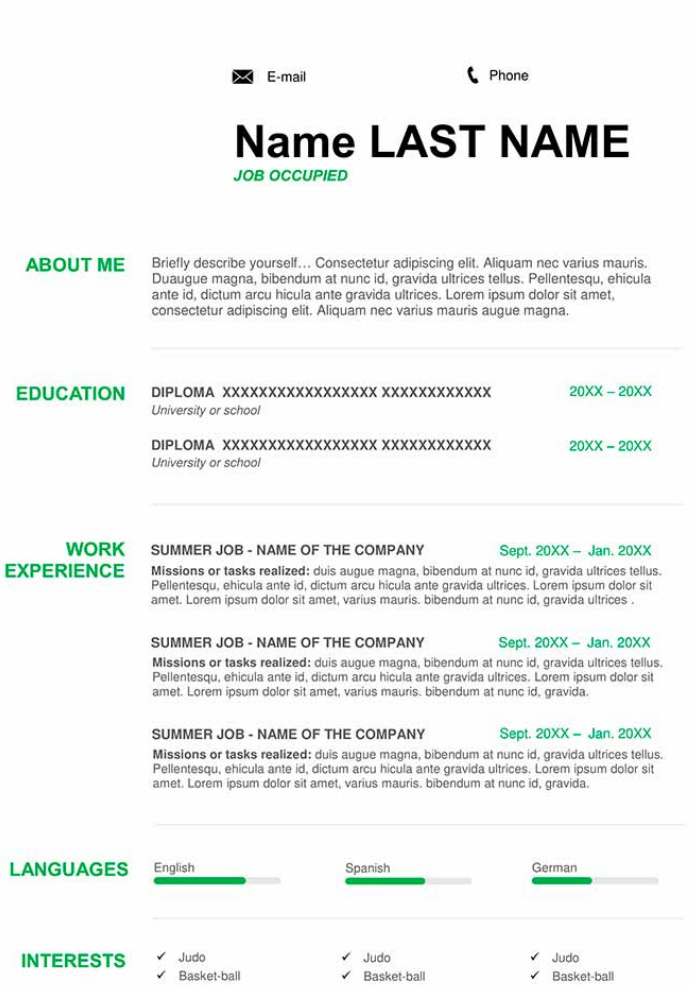
A primer on Ethics

Ethics doesn't need to be abstract

CV Assessment

1. You send a CV for a job.
2. The company automatically pre-filters candida
3. Do you pass?
4. 🔴🟢

Interview Assessment



E-mail Phone

Name LAST NAME

JOB OCCUPIED

ABOUT ME

Briefly describe yourself… Consectetur adipiscing elit. Aliquam nec varius mauris. Duaugue magna, bibendum at nunc id, gravida ultrices tellus. Pellentesqu, ehicula ante id, dictum arcu hicula ante gravida ultrices. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam nec varius mauris augue magna.

EDUCATION

DIPLOMA XXXXXXXXXXXXXXXX XXXXXXXXXXXX 20XX – 20XX
University or school

DIPLOMA XXXXXXXXXXXXXXXX XXXXXXXXXXXX 20XX – 20XX
University or school

WORK EXPERIENCE

SUMMER JOB - NAME OF THE COMPANY Sept. 20XX – Jan. 20XX
Missions or tasks realized: duis augue magna, bibendum at nunc id, gravida ultrices tellus. Pellentesqu, ehicula ante id, dictum arcu hicula ante gravida ultrices. Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, varius mauris. bibendum at nunc id, gravida ultrices .

SUMMER JOB - NAME OF THE COMPANY Sept. 20XX – Jan. 20XX
Missions or tasks realized: duis augue magna, bibendum at nunc id, gravida ultrices tellus. Pellentesqu, ehicula ante id, dictum arcu hicula ante gravida ultrices. Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, varius mauris. bibendum at nunc id, gravida.

SUMMER JOB - NAME OF THE COMPANY Sept. 20XX – Jan. 20XX
Missions or tasks realized: duis augue magna, bibendum at nunc id, gravida ultrices tellus. Pellentesqu, ehicula ante id, dictum arcu hicula ante gravida ultrices. Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, varius mauris. bibendum at nunc id, gravida.

LANGUAGES

English Spanish German

INTERESTS

✓ Judo ✓ Judo ✓ Judo
✓ Basket-ball ✓ Basket-ball ✓ Basket-ball

Don’t worry about the future …

… worry about the present.

- AI is persuasive in everyday’s life
- AI is extremely capable
- AI is not so smart

💡 what is going on behind the scene?

An (R)evolution of NLP

From Bag of Words to Embeddings

Putting Words into Computers *(Smith 2020; **Church2022?**)*

- from **coarse, static** to **fine, contextual** meaning
- how to measure similarity
  - *string-based*
  - *syntactic (e.g., part-of-speech)*
  - *semantic (e.g., animate)*
  - *embedding*
- from counting to learning representations
  - *end2end deep learning*

Bag of Words

- word as arbitrary, discrete numbers

  `King = 1, Queen = 2, Man = 3, Woman = 4`

- intrinsic meaning

  derived from data

  defined by experts

- how are these words similar?

Corpus Representation

Collection of Documents

1. `NLP is great. I love NLP.`
2. `I understand NLP.`
3. `NLP, NLP, NLP.`

Document Term Matrix

	`NLP`	`I`	`is`	***term***
Doc 1	2	1	1	...
Doc 2	1	1	0	...
Doc 3	3	0	0	...
Doc ID	...	...	...	*term frequency*

«You shall know a word by the company it keeps!»

Firth (1957)

"I hate ____, but the others eat anything."

Word Embeddings

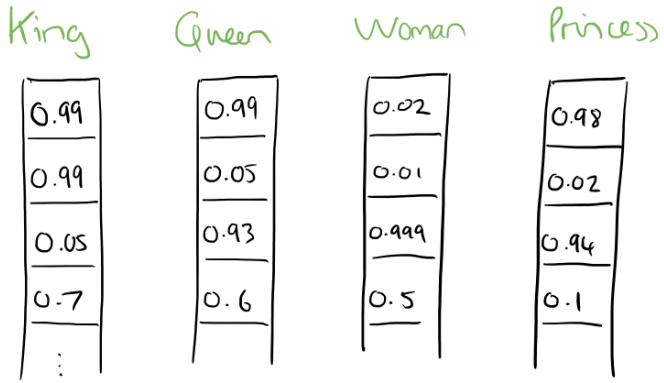
word2vec *(Mikolov et al. 2013)*

- words as continuous vectors

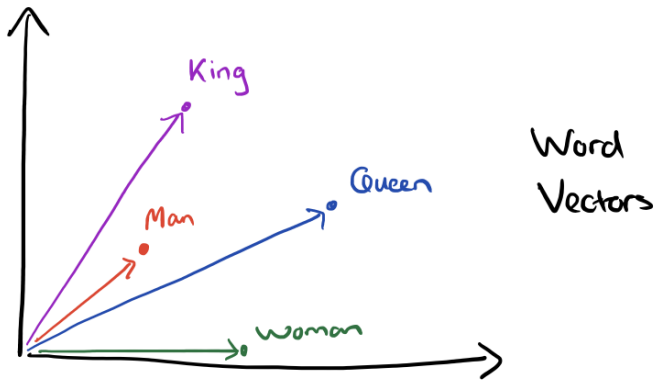
  accounting for similarity between words

- semantic similarity

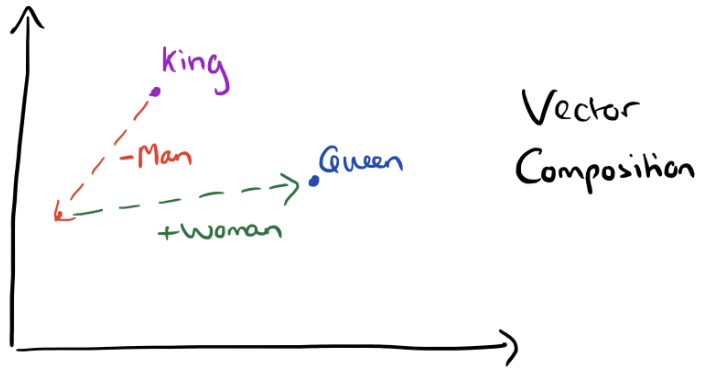
```
King - Man + Woman = Queen
```



Single continuous vector per word *(Colyer 2016)*



Words as points in a semantic space *(Colyer 2016)*



Doing arithmetics with words *(Colyer 2016)*

Contextualized Word Embeddings

BERT *(Devlin et al. 2019)*

- recontextualize static word embedding
  - *different embedding in different contexts*
  - *accounting for ambiguity (e.g.,* `bank`*)*
- acquire linguistic knowledge from language models (LM)
  - *LM predict next/missing word*
  - *pre-trained on massive data (> 1 billion words)*

💥 embeddings mark a new era of NLP

(Cultural) Associations in Data

- «___ becomes a doctor.»
- «___ takes care of the children.»

- «Doing illegal things is bad.»
- «A recent study by the Swiss state secretariat for migrants (SEM) estimates that there were 76,000 illegal immigrants.»

🤔 Are migrants somehow bad, then?

Word vectors are biased ...

... because ~~our data is~~ we are biased. *(Bender2022?)*

Two Sides of the AI Coin

Explaining vs. Solving

- understanding matters in **science**: **data analysis**
- automating matters in **business**: **applied AI**

Biased Data and beyond

Data = Digital Traces = Social Artifacts

- collecting, curating, preserving traces
- data is a **tool** to refine questions **rather than reflection** of the world

Data vs. Capta

*«Differences in the etymological roots of the terms data and capta make the distinction between constructivist and realist approaches clear. Capta is **“taken”** actively while data is assumed to be a **“given”** able to be recorded and observed.»*

Imperfect Data: A Tail of Bias

- social bias

  view from somewhere, stereotypes

- data/archive holes

  lost, uncollected

- corpus curation

  supposition that key-word indicates topic

- noise in data

  OCR errors, inconsistent spelling, non-content

👉 think about the data and mitigate issues

«Raw data is an oxymoron.» *Gitelman (2013)*

Mind your Data

- Who has a voice in your data?
  social context
- bigger is not necessarily better
  more vs. more diverse data
- clean your data thoroughly
  noisy vs. clean data

DATA HUMANISM

SMALL ~~big~~ data

data bandwith QUALITY

IMPERFECT ~~infallible~~ data

SUBJECTIVE ~~impartial~~ data

INSPIRING ~~descriptive~~ data

SERENDIPITOUS ~~predictive~~ data

data conventions POSSIBILITIES

data to simplify complexity / DEPICT

data processing DRAWING

data driven design

SPEND ~~save~~ time with data

data is numbers PEOPLE

data will make us more efficient HUMAN.

@giorgialupi

Fair is a Fad

- *Fair* and *good* – but to whom? *(Kalluri 2020)*
- lacking democratic legitimacy
- debiasing data to avoid regulation
- looking beyond data
  - *invading privacy*
  - *economic monopolies*
  - *(unpaid) AI-trainers and click-workers*
  - *environmental costs*

«Don’t ask if artificial intelligence is good or fair, ask how it shifts power.»

Kalluri (2020)


Anatomy of an AI system
An anatomical case study of the Amazon echo as a artificial intelligence system made of human labor
Human operator
Domestic infrastructure
Internet infrastructure
Internet Platforms & Services
AI Training
Data Preparation and Labeling
Income distribution
Distributors
Assemblers
Transportation
Component manufacturers
Mines
Elements
Geological process
Abandoned devices
Collecting
Shipping
Disposing
Geological process
Quantification of nature
Data exploitation
Anatomy of AI (Crawford and Joler 2018)
illustrated by the Amazon Echo

Data represents real life.

Don’t be a fool. Be wise, think twice.


Questions?

References

Colyer, Adrian. 2016. “The Amazing Power of Word Vectors.” the morning paper. 2016. https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/.

Crawford, Kate, and Vladan Joler. 2018. “Anatomy of an AI System.” Anatomy of an AI System. 2018. http://www.anatomyof.ai.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” May 24, 2019. http://arxiv.org/abs/1810.04805.

Drucker, Johanna. 2011. “Humanities Approaches to Graphical Display.” *Digital Humanities Quarterly* 5 (1). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Firth, John R. 1957. “A Synopsis of Linguistic Theory, 1930-1955.” In *Studies in Linguistic Analysis: Special Volume of the Philological Society*, edited by John R. Firth, 1–32. Oxford: Blackwell. http://ci.nii.ac.jp/naid/10020680394/.

Gitelman, Lisa. 2013. *Raw Data Is an Oxymoron*. Cambridge: MIT.

Kalluri, Pratyusha. 2020. “Don’t Ask If Artificial Intelligence Is Good or Fair, Ask How It Shifts Power.” *Nature* 583 (7815, 7815): 169–69. https://doi.org/10.1038/d41586-020-02003-2.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. “Distributed Representations of Words and Phrases and Their Compositionality.” In *Advances in Neural Information Processing Systems*, 3111–19.

Smith, Noah A. 2020. “Contextual Word Representations: Putting Words into Computers.” *Communications of the ACM* 63 (6): 66–74. https://doi.org/10.1145/3347145.