open space

Feel free to approach us in case of questions…
(microphone or chat)

«Critical Social Media Analysis using Mixed Methods»

Language Models

Michael Tebbe, Dr. Simon David Hirsbrunner

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

Session III, 19 Nov 2020

Recap last session

Epistemological precautions

- Browsers, Accounts, APIs

Ethical and legal considerations in SMA

- open software, not-open datasets

Data collection

- YouTube Data Tools
- YouTube Data API

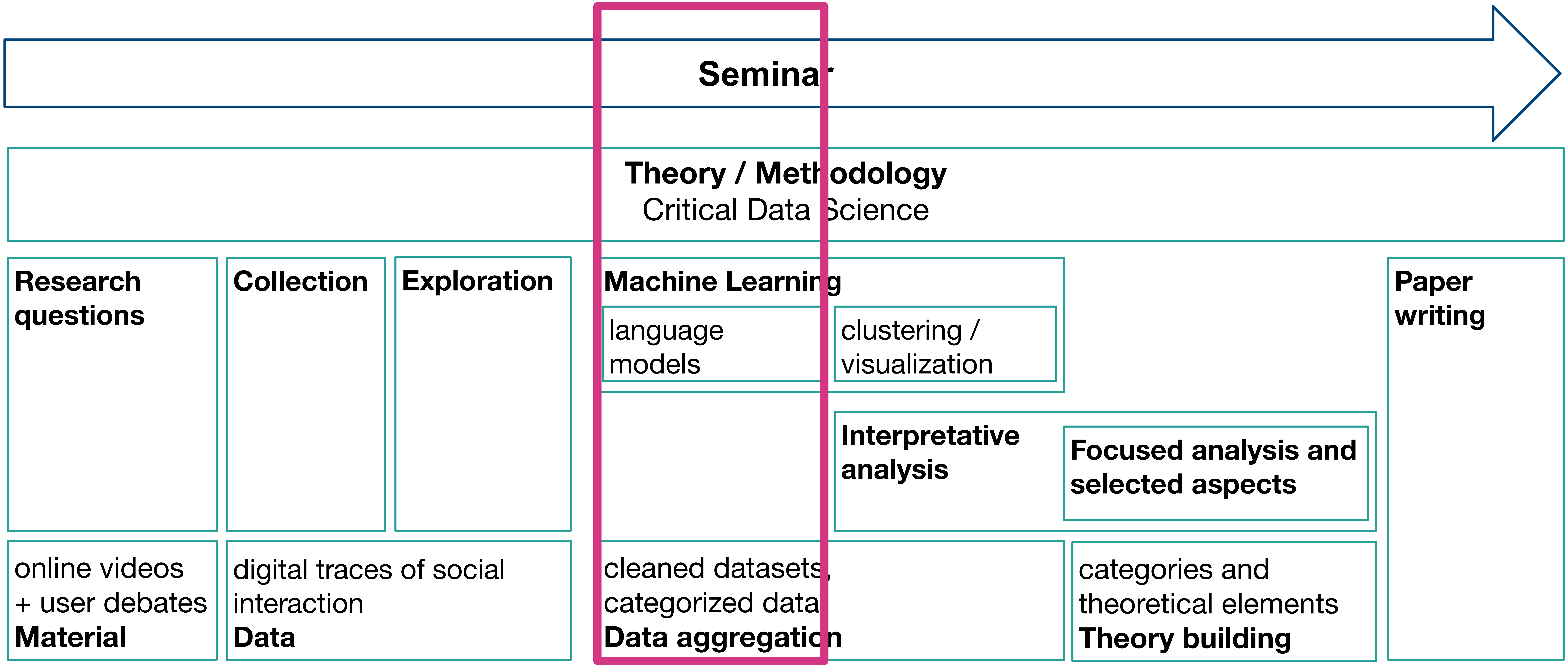
Data exploration

Assignments

Plan for today

- Collaborative collection of ideas and meeting peers
- Language Models
- (Short break)
- Sentence Embeddings with Universal Sentence Encoder
- Assignments

Seminar progress / today



Seminar

Theory / Methodology
Critical Data Science

Research questions

Collection

Exploration

Machine Learning

language models

clustering / visualization

Interpretative analysis

Focused analysis and selected aspects

Paper writing

online videos
+ user debates
Material

digital traces of social interaction
Data

cleaned datasets, categorized data
Data aggregation

categories and theoretical elements
Theory building

Collaborative brainstorming and meeting peers

Go to: https://flinga.fi/s/FL249B5

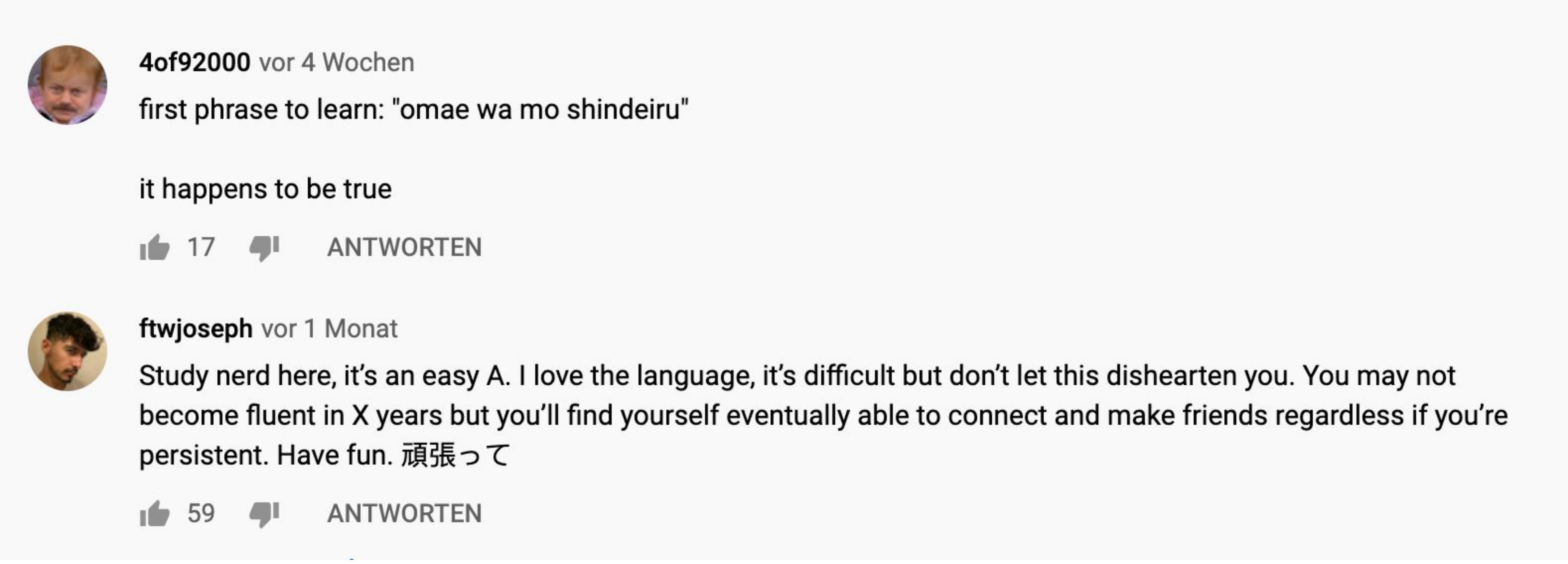
Flinga is a tool for collaborative brainstorming and visualization.

Tasks for collaborative collection of ideas and networking

- Use square post-it’s to write prospective topics for further investigation and locate them on the canvas.
  - You can frame as question, topic or approach.
  - Positioning of post-its according to Mapping Controversy modes of inquiry
- Use the people symbol to create an avatar for you and add your name. Position yourself near one topic you find interesting
- If several people gather around one topic, we create a breakout room in WebEx, so you can discuss further and meet your peers. Instructors will drag a circle with the specification of the breakout room near your group.
- Join the indicated breakout room on Webex and discuss your subject.
  - Exchange email-addresses if you would like to collaborate in the future (e.g. for assignments and the seminar project)
  - Create an etherpad to document your discussions: https://pad.spline.inf.fu-berlin.de/
  - Post the link of the etherpad on GitHub

Info: groups for seminar projects

Language Modeling



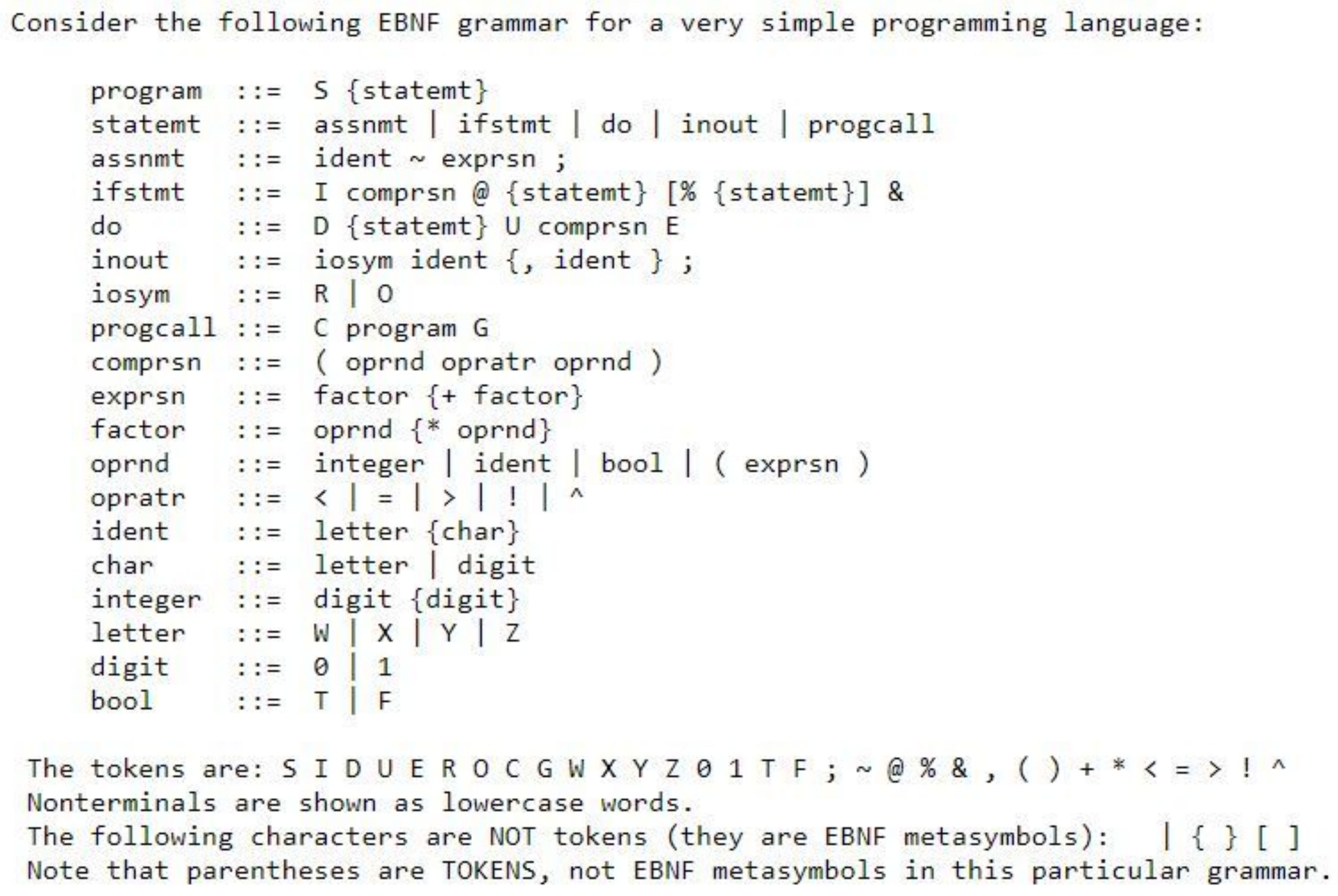
4of92000 vor 4 Wochen

first phrase to learn: "omae wa mo shindeiru"

it happens to be true

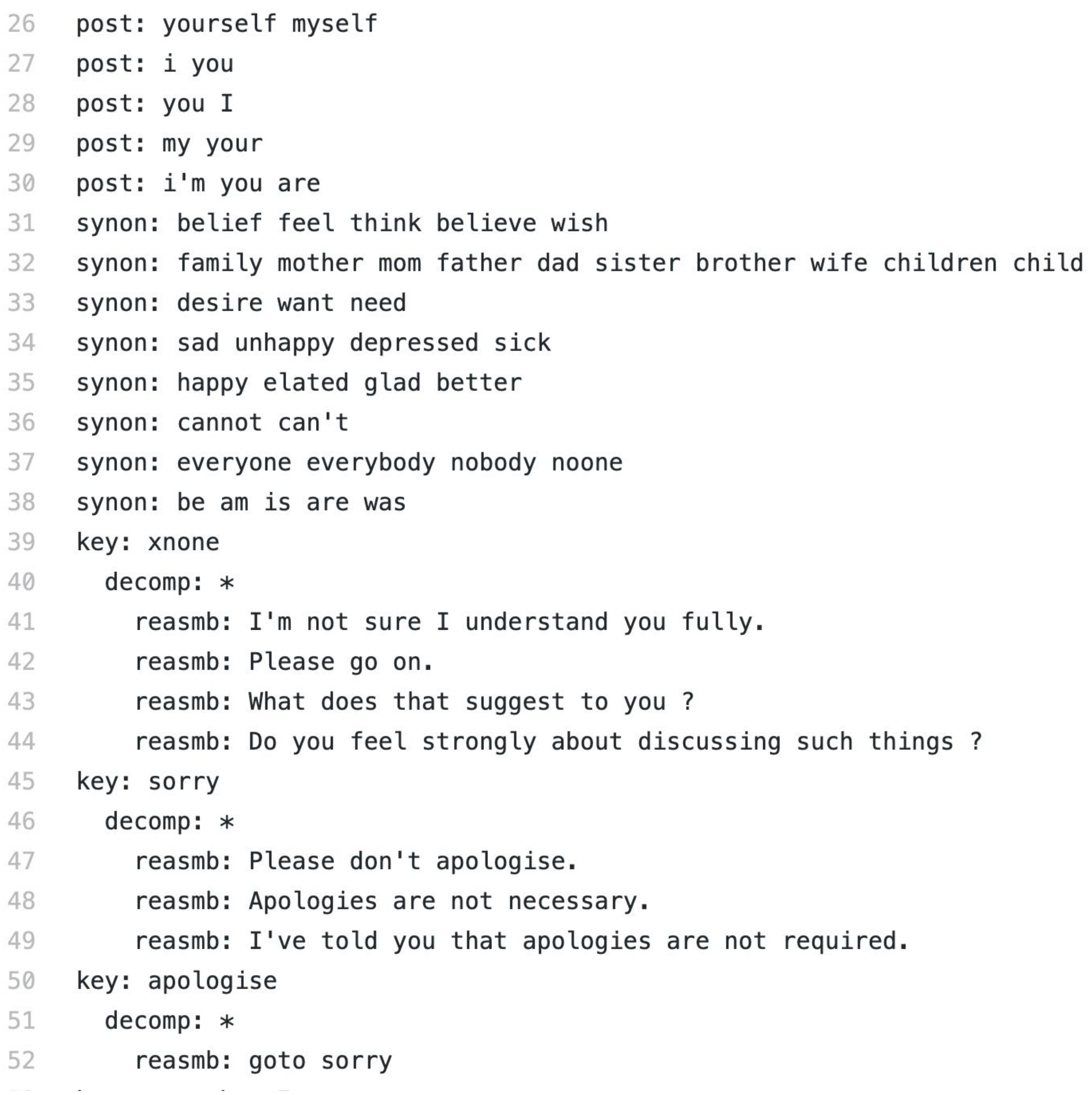
ftwjoseph vor 1 Monat

Study nerd here, it's an easy A. I love the language, it's difficult but don't let this dishearten you. You may not become fluent in X years but you'll find yourself eventually able to connect and make friends regardless if you're persistent. Have fun. 頑張って

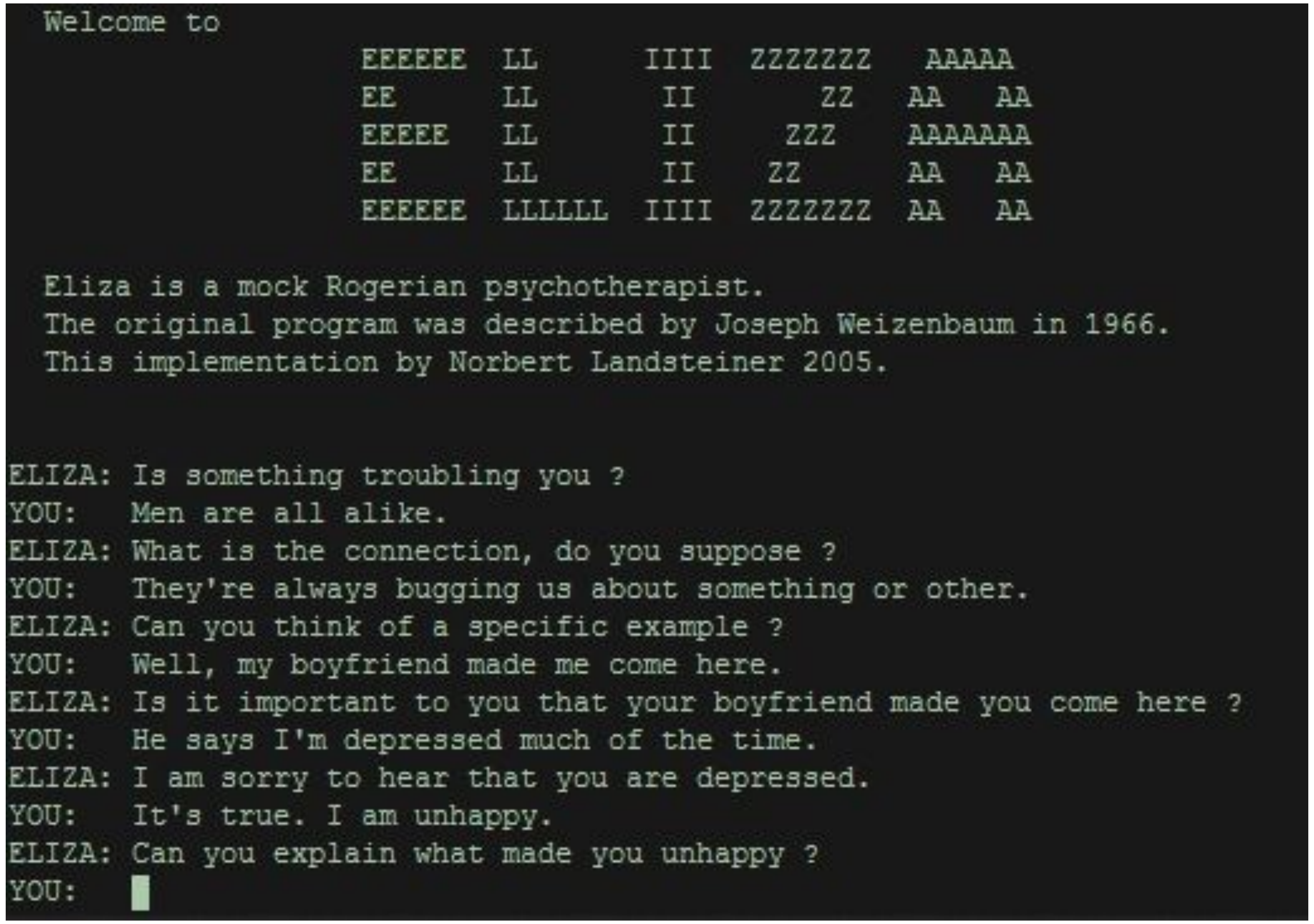


```
Consider the following EBNF grammar for a very simple programming language:

     program  ::=  S {statemt}
     statemt  ::=  assnmt | ifstmt | do | inout | progcall
     assnmt   ::=  ident ~ exprsn ;
     ifstmt   ::=  I comprsn @ {statemt} [% {statemt}] &
     do       ::=  D {statemt} U comprsn E
     inout    ::=  iosym ident {, ident } ;
     iosym    ::=  R | O
     progcall ::=  C program G
     comprsn  ::=  ( oprnd opratr oprnd )
     exprsn   ::=  factor {+ factor}
     factor   ::=  oprnd {* oprnd}
     oprnd    ::=  integer | ident | bool | ( exprsn )
     opratr   ::=  < | = | > | ! | ^
     ident    ::=  letter {char}
     char     ::=  letter | digit
     integer  ::=  digit {digit}
     letter   ::=  W | X | Y | Z
     digit    ::=  0 | 1
     bool     ::=  T | F

 The tokens are: S I D U E R O C G W X Y Z 0 1 T F ; ~ @ % & , ( ) + * < = > ! ^
 Nonterminals are shown as lowercase words.
 The following characters are NOT tokens (they are EBNF metasymbols):   | { } [ ]
 Note that parentheses are TOKENS, not EBNF metasymbols in this particular grammar.
```

Language Modeling - Symbolic NLP



```
post: yourself myself
post: i you
post: you I
post: my your
post: i'm you are
synon: belief feel think believe wish
synon: family mother mom father dad sister brother wife children child
synon: desire want need
synon: sad unhappy depressed sick
synon: happy elated glad better
synon: cannot can't
synon: everyone everybody nobody noone
synon: be am is are was
key: xnone
  decomp: *
    reasmb: I'm not sure I understand you fully.
    reasmb: Please go on.
    reasmb: What does that suggest to you ?
    reasmb: Do you feel strongly about discussing such things ?
key: sorry
  decomp: *
    reasmb: Please don't apologise.
    reasmb: Apologies are not necessary.
    reasmb: I've told you that apologies are not required.
key: apologise
  decomp: *
    reasmb: goto sorry
```

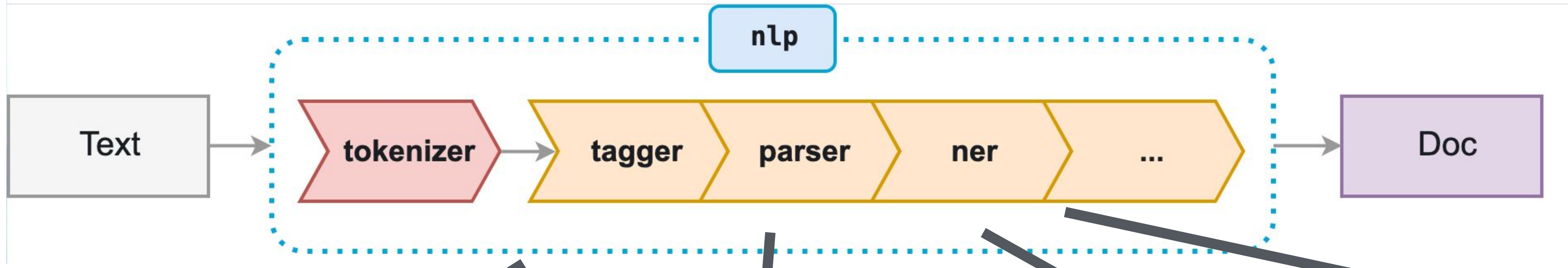


```
  Welcome to
                    EEEEEE  LL      IIII  ZZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   LL       II      ZZZ   AAAAAAA
                    EE      LL       II     ZZ     AA   AA
                    EEEEEE  LLLLLL  IIII  ZZZZZZZ  AA   AA

  Eliza is a mock Rogerian psychotherapist.
  The original program was described by Joseph Weizenbaum in 1966.
  This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:   █
```

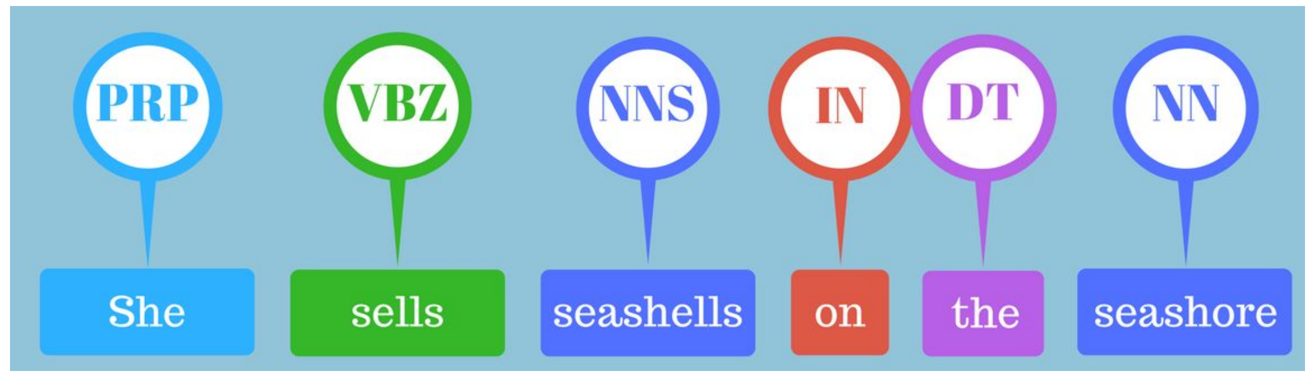
https://github.com/codeanticode/eliza/blob/master/src/codeanticode/eliza/Eliza.java

Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. <i>Commun. ACM</i> 9, 1 (Jan. 1966), 36–45. DOI:https://doi.org/10.1145/365153.365168

Language Modeling - Statistical NLP

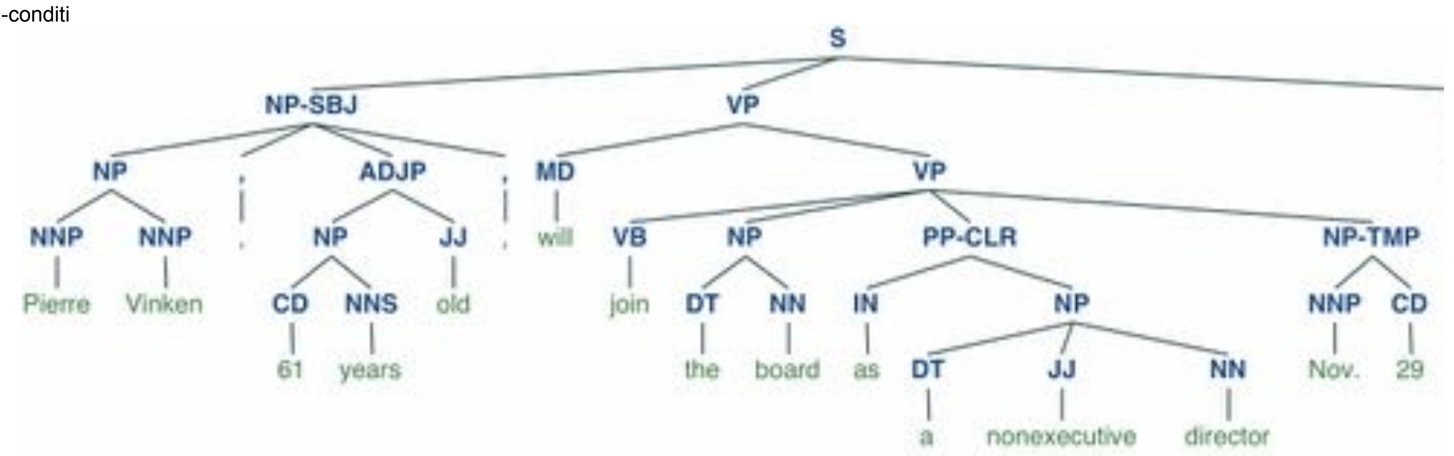


https://spacy.io/usage/processing-pipelines



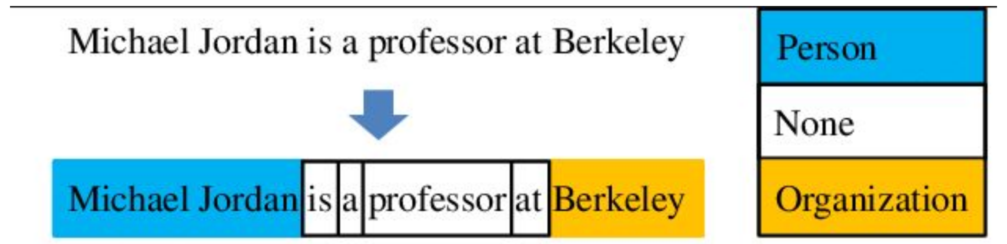
https://medium.com/analytics-vidhya/pos-tagging-using-conditional-random-fields-92077e5eaa31

POS-Tagging



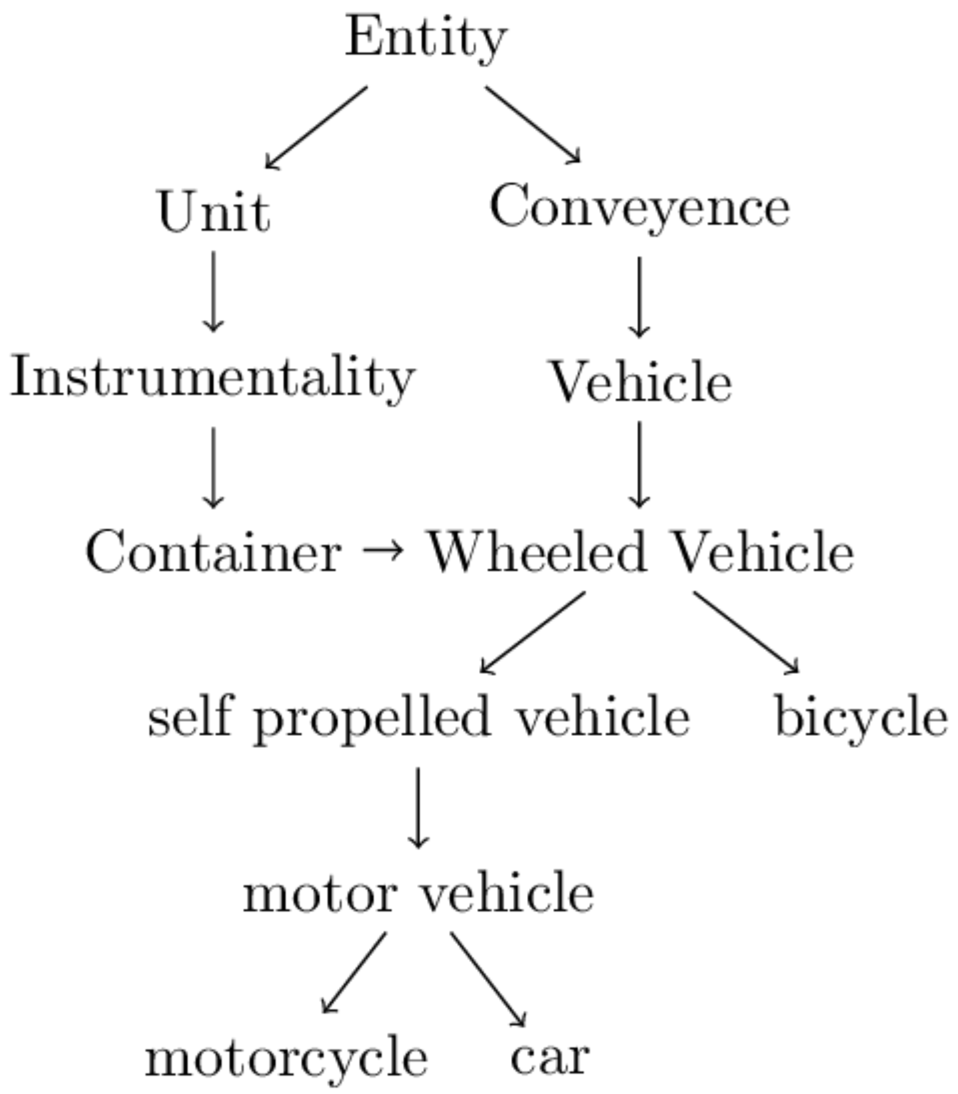
Parse Tree

http://www.nltk.org/



Named Entity Recognition

Liu, Yijia & Che, Wanxiang & Guo, Jiang & Qin, Bing & Liu, Ting. (2016). Exploring Segment Representations for Neural Segmentation Models.



Semantic Web

Langugage Modeling - Neural NLP

sparse vector with one 1 and many zeros:

[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0]

- Dimensionality: vocabulary size e.g.:
  20K (speech) – 50K (PTB) – 500K (large corpus)
  Hotel [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 . . . 0 0 0 1 0 0 0]
  Motel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 . . . 0 0 0 0 0 0 0]

https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/

One-hot encoding

-> Vectorized Representation of documents

Problem: very sparse, hard to determine similarity (Curse of Dimensionality)

Mikolov, Chen, Corrado & Dean (2013): Efficient Estimation of Word Representations in Vector Space

Harris, Z. (1954). "Distributional structure". *Word*. **10** (23): 146–162. doi:10.1080/00437956.1954.11659520.

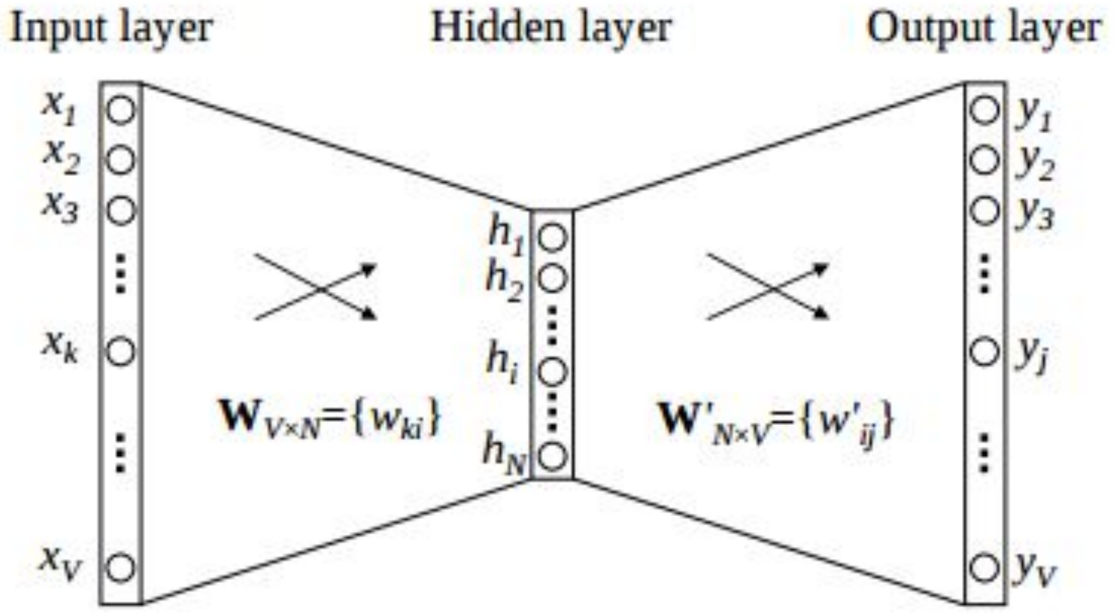
Firth, J.R. (1957). "A synopsis of linguistic theory 1930-1955"

🟦 : Center Word
🟥 : Context Word

c=0 The cute cat jumps over the lazy dog.

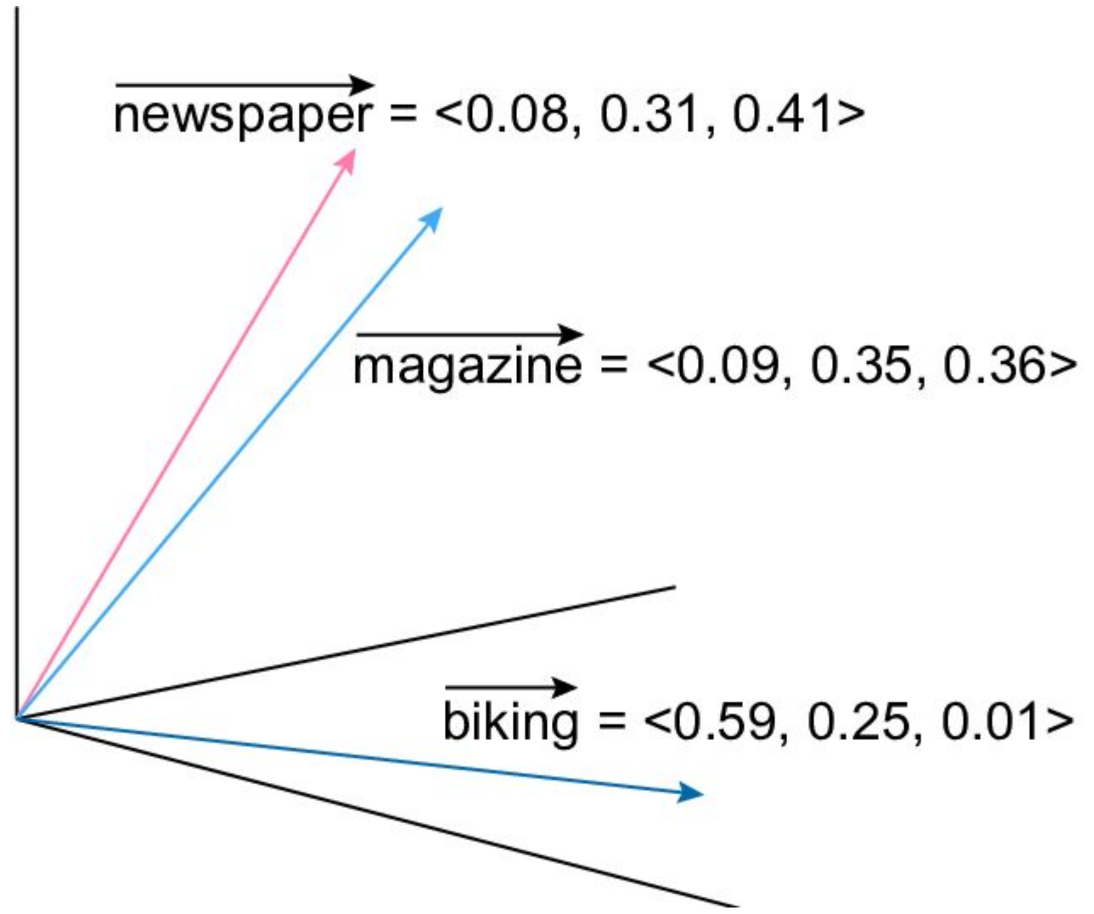
c=1 The cute cat jumps over the lazy dog.

c=2 The cute cat jumps over the lazy dog.



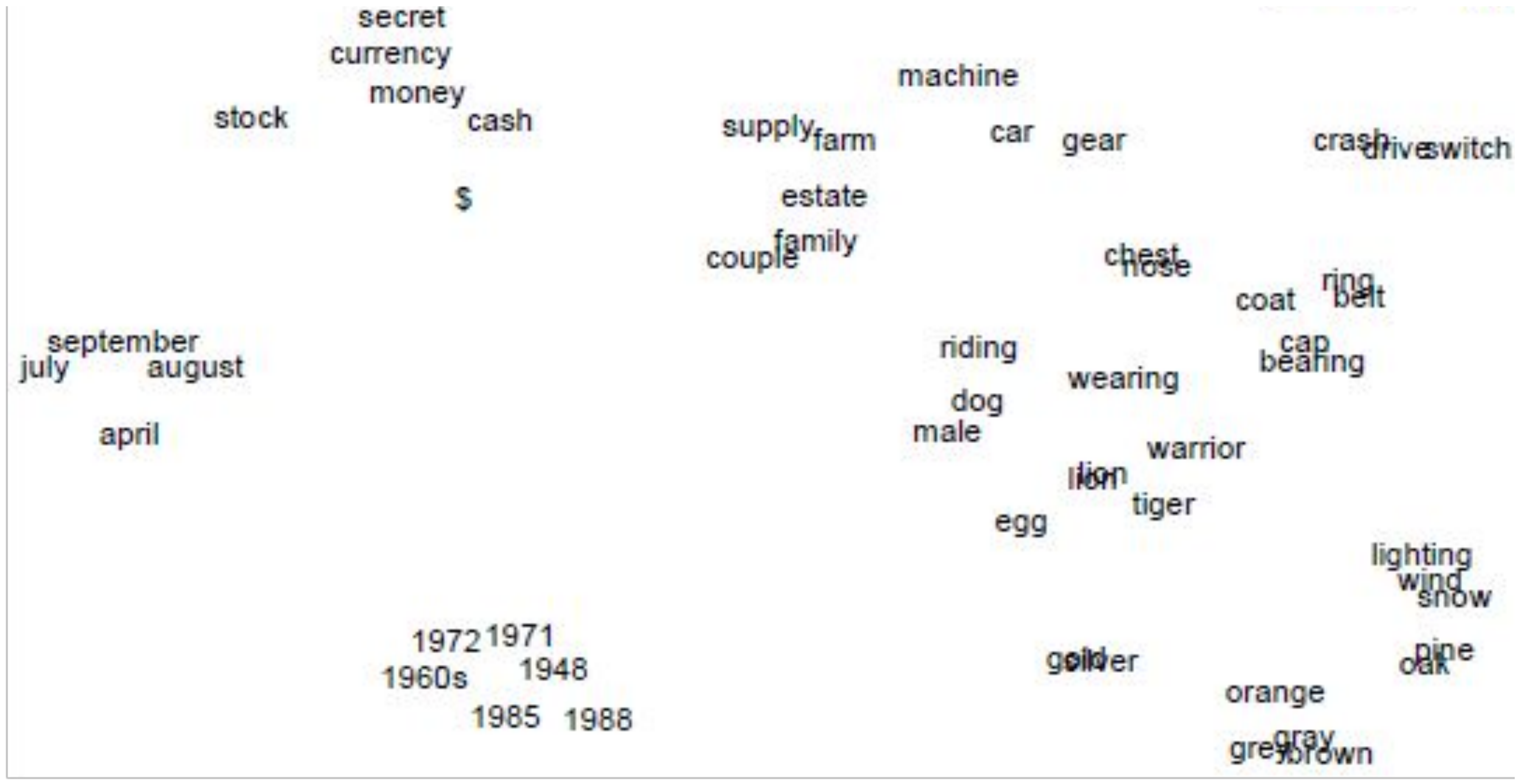
Word2Vec

“You shall know a word by the company it keeps”
(J.R.Firth, 1957)



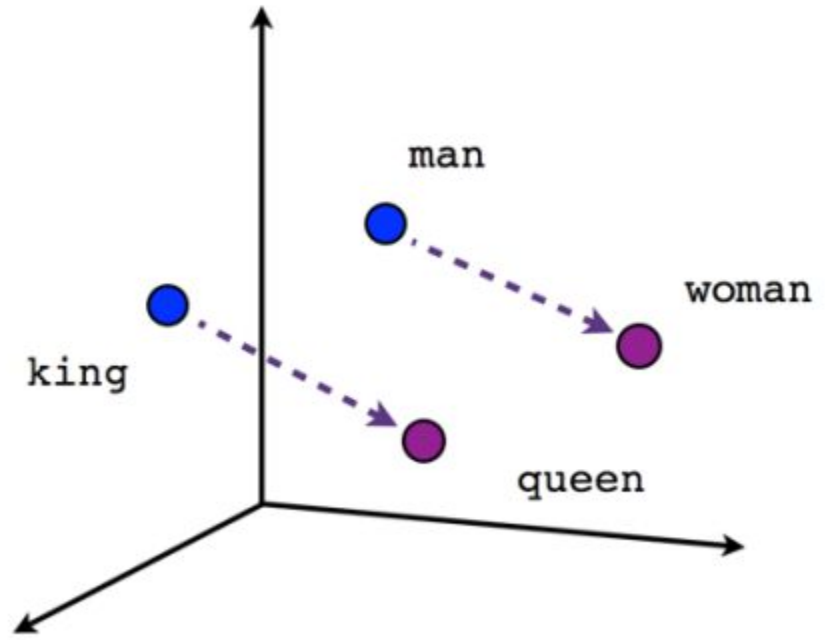
http://mbenhaddou.com/2019/12/14/word2vec-concepts-from-scratch/

Langugage Modeling - Neural NLP



Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1 (ACL '12). Association for Computational Linguistics, USA, 873–882.

Problem: Polysemy

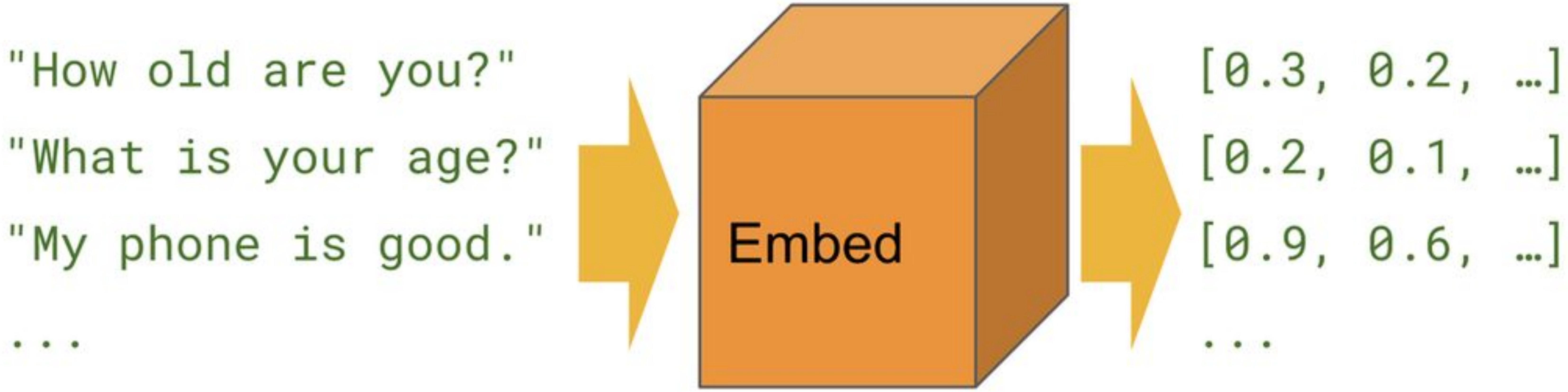


king - man + woman = queen

https://towardsdatascience.com/creating-word-embeddings-coding-the-word2vec-algorithm-in-python-using-deep-learning-b337d0ba17a8

Short break: 5 Minutes

Universal Sentence encoder



Universal Sentence encoder - Demo

Sanders, Abraham, Rachael White, Lauren Severson, Rufeng Ma, Richard McQueen, Haniel Campos Alcanatara Paulo, Yucheng Zhang, John S Erickson, und Kristin P Bennett. „Unmasking the Conversation on Masks: Natural Language Processing for Topical Sentiment Analysis of COVID-19 Twitter Discourse". Preprint. Health Informatics, 1. September 2020. https://doi.org/10.1101/2020.08.28.20183863.

Universal Sentence encoder -Input Data

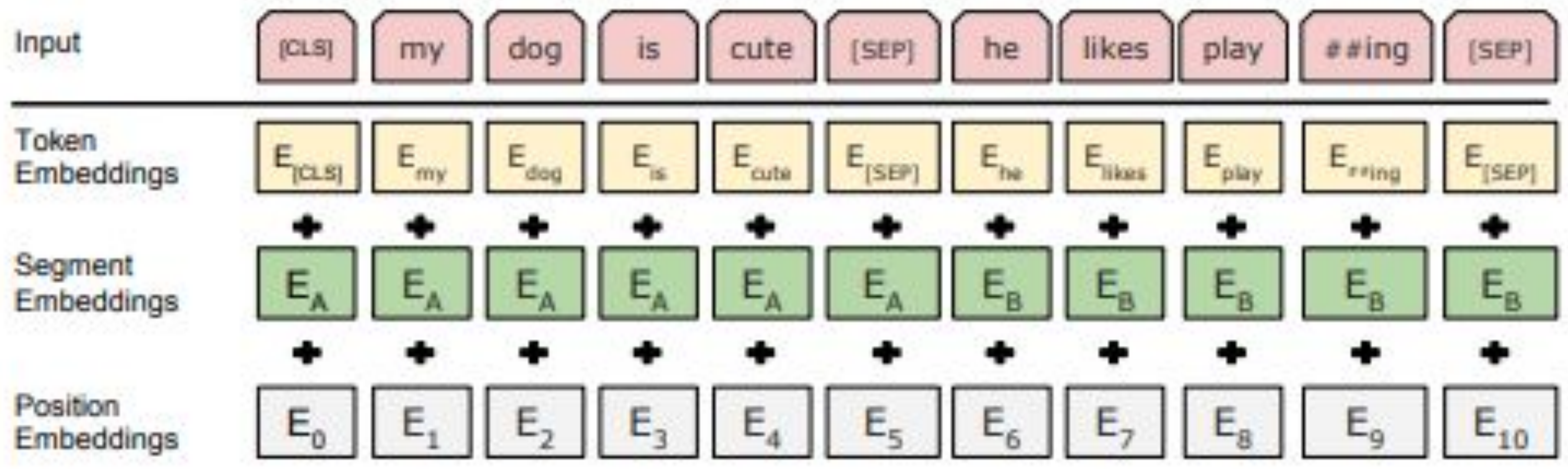


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Universal Sentence encoder - Versions

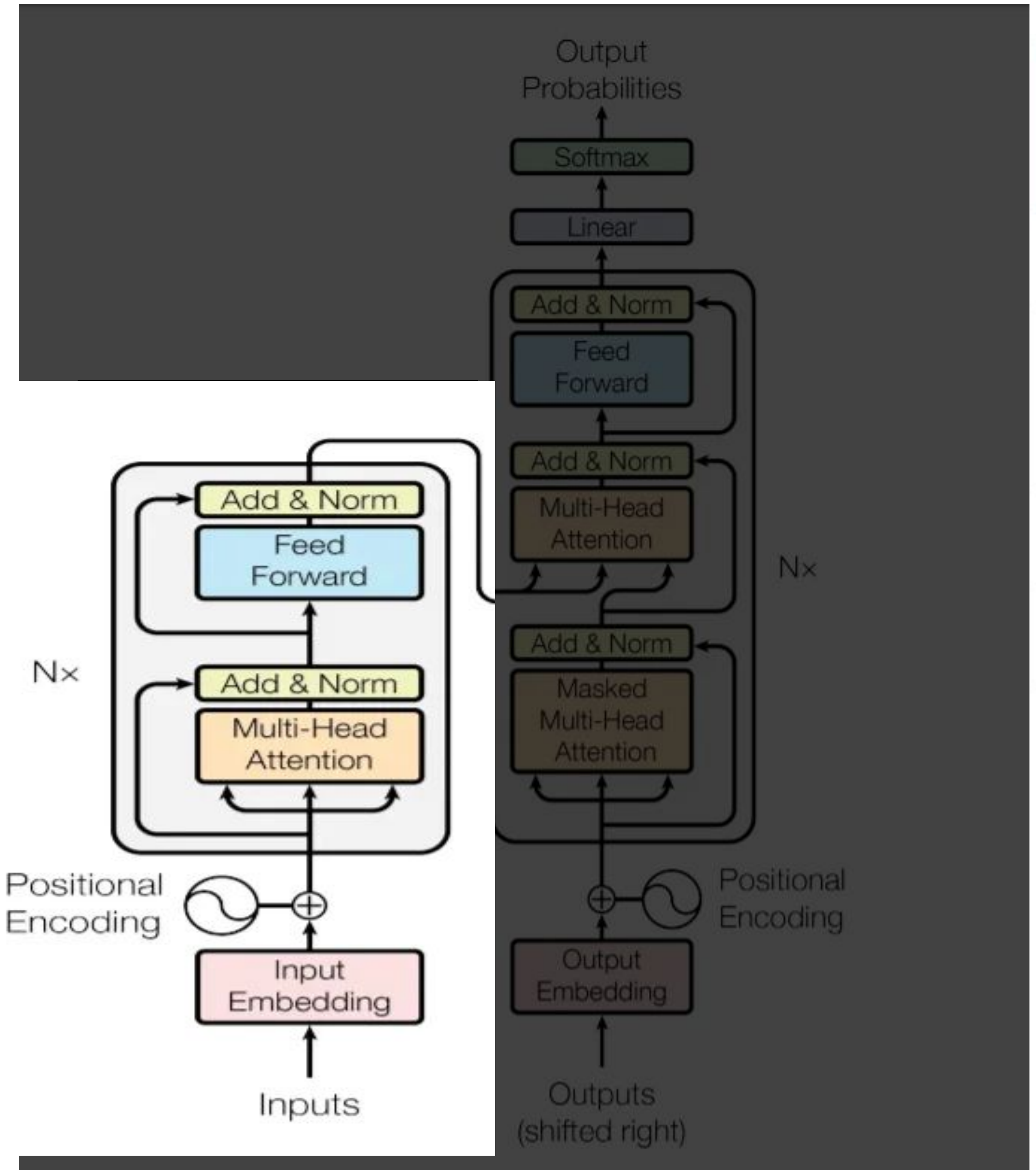
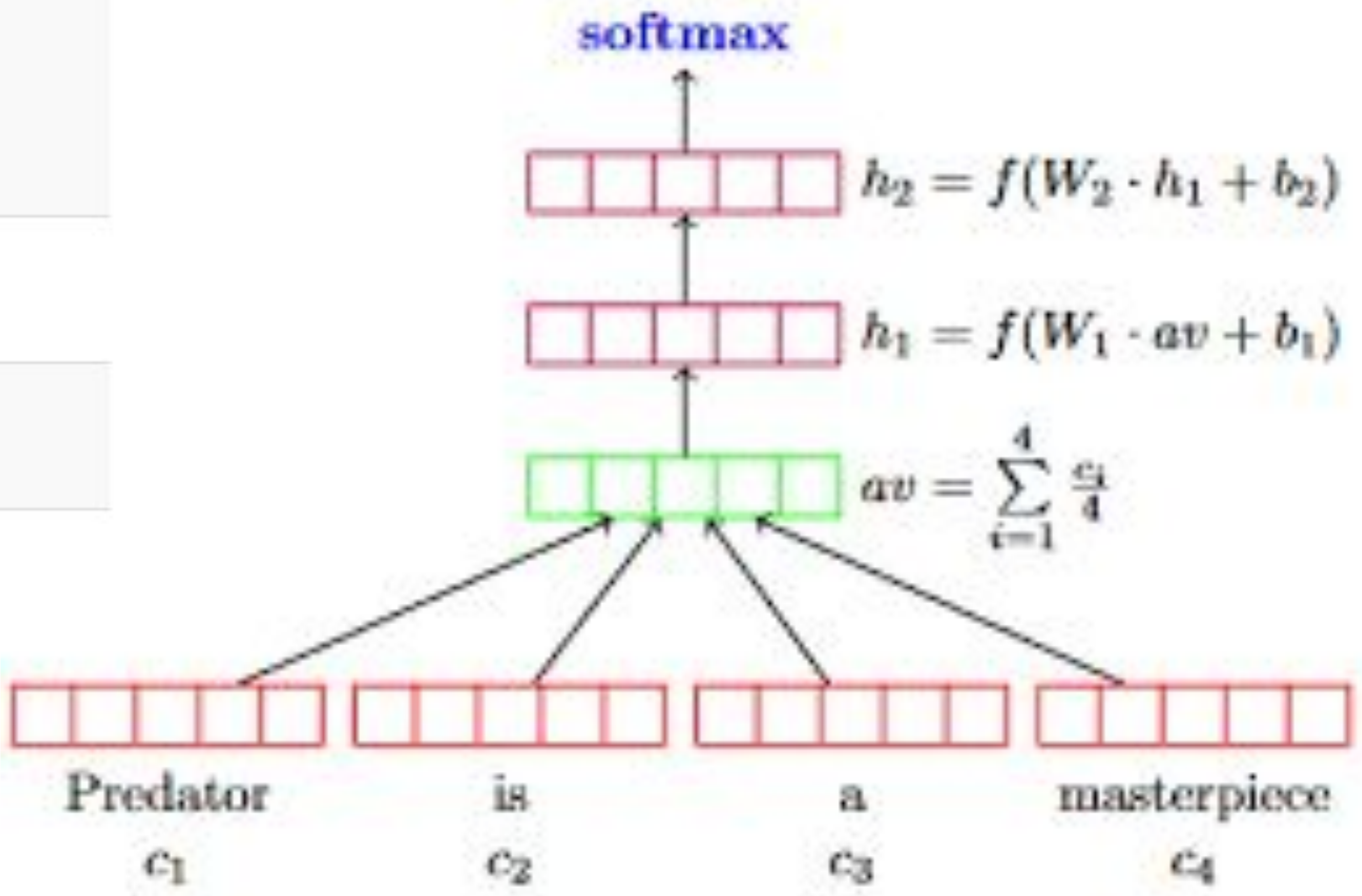


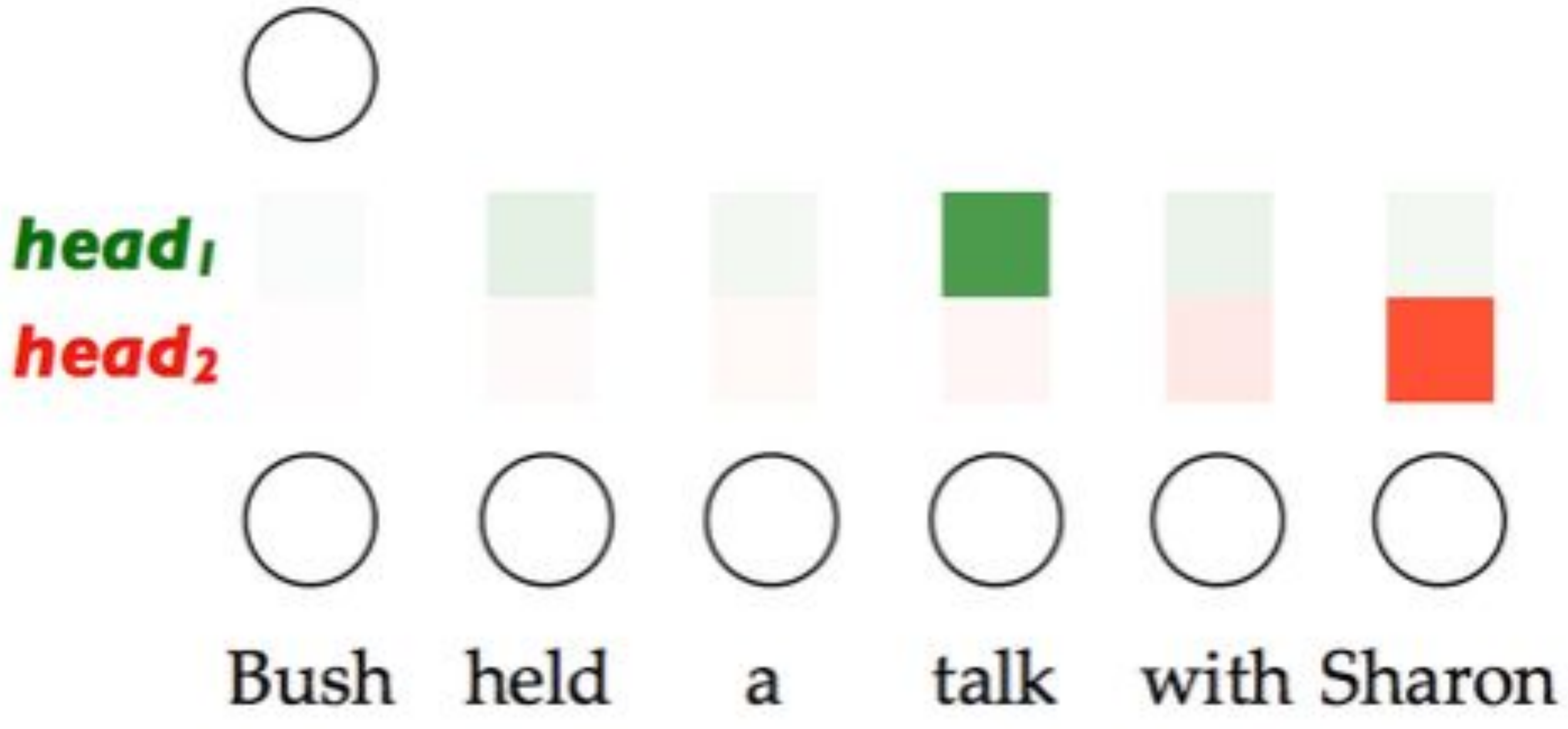
Figure 1: The Transformer - model architecture.

	Transformer model	Deep Averaging Network (DAN) model
Vector Length	512	512
Encoding time with sentence length	Non-Linear	Linear
Memory usage	High	Medium
Accuracy	Very High	High

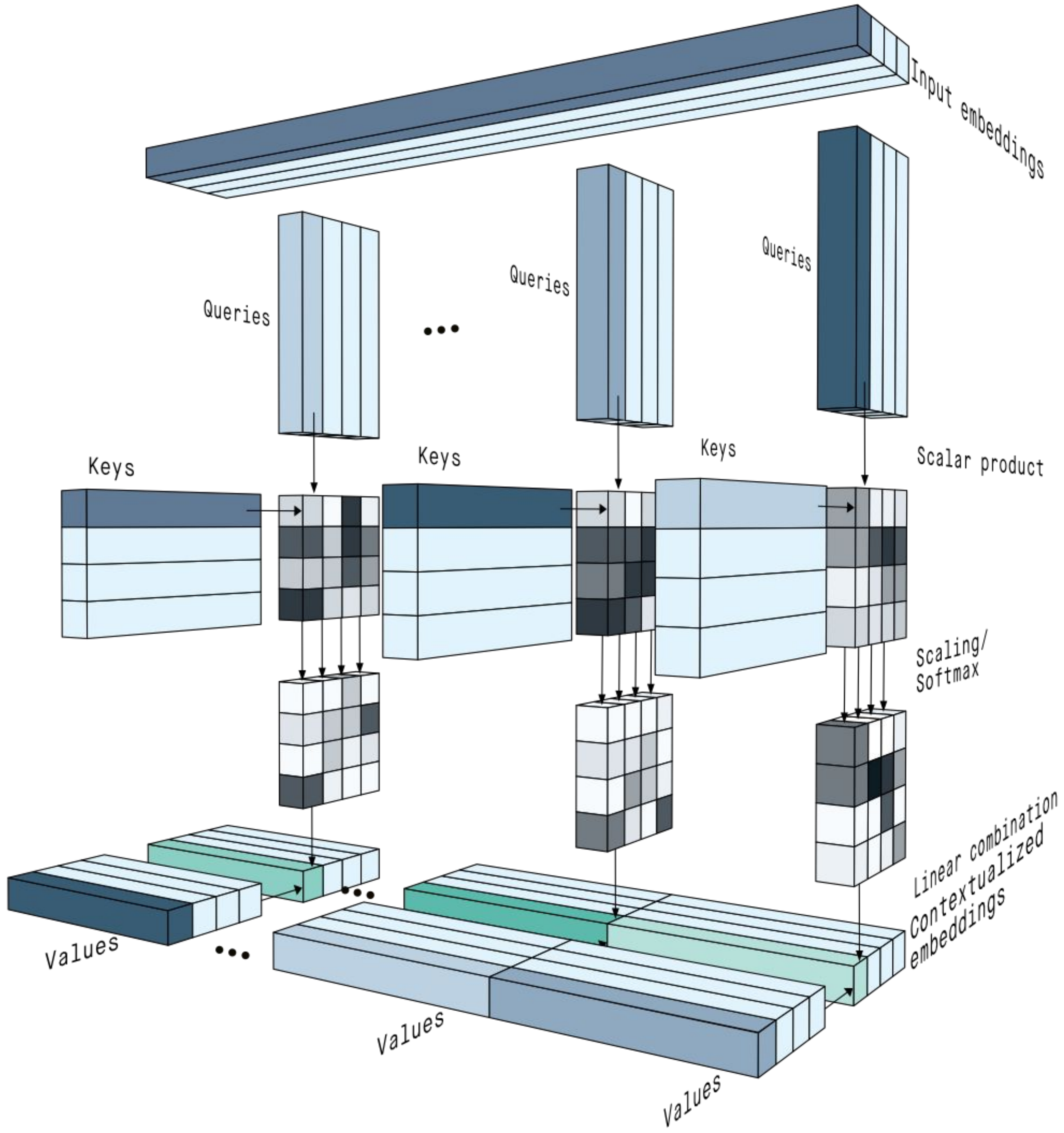


softmax
$h_2 = f(W_2 \cdot h_1 + b_2)$
$h_1 = f(W_1 \cdot av + b_1)$
$av = \sum_{i=1}^{4} \frac{c_i}{4}$
Predator
c_1
is
c_2
a
c_3
masterpiece
c_4

Universal Sentence encoder - Attention



Attention



Universal Sentence encoder - What does the model encode?

- Linguistic Structure [1]:
  - word morphology
  - part-of-speech information
  - lexical semantics
  - non-local syntactic and semantic dependencies
- Morality [2]
- Social Bias [3]

[1] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the Linguistic Representational Power of Neural Machine Translation Models. Comput. Linguist. 46, 1 (March 2020), 1–52. DOI:https://doi.org/10.1162/coli_a_00367

[2] Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics Derived Automatically from Language Corpora Contain Human-like Moral Choices. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). Association for Computing Machinery, New York, NY, USA, 37–44. DOI:https://doi.org/10.1145/3306618.3314267

[3] Cer, D.M., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder. ArXiv, abs/1803.11175.

Universal Sentence encoder - Use Cases

- Translation [1], Semantic retrieval, Semantic similarity [1], Outlier detection [2]
- Detecting depression [3]
- Fact checking [4]
- …

[1] Cer, D.M., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder. ArXiv, abs/1803.11175.

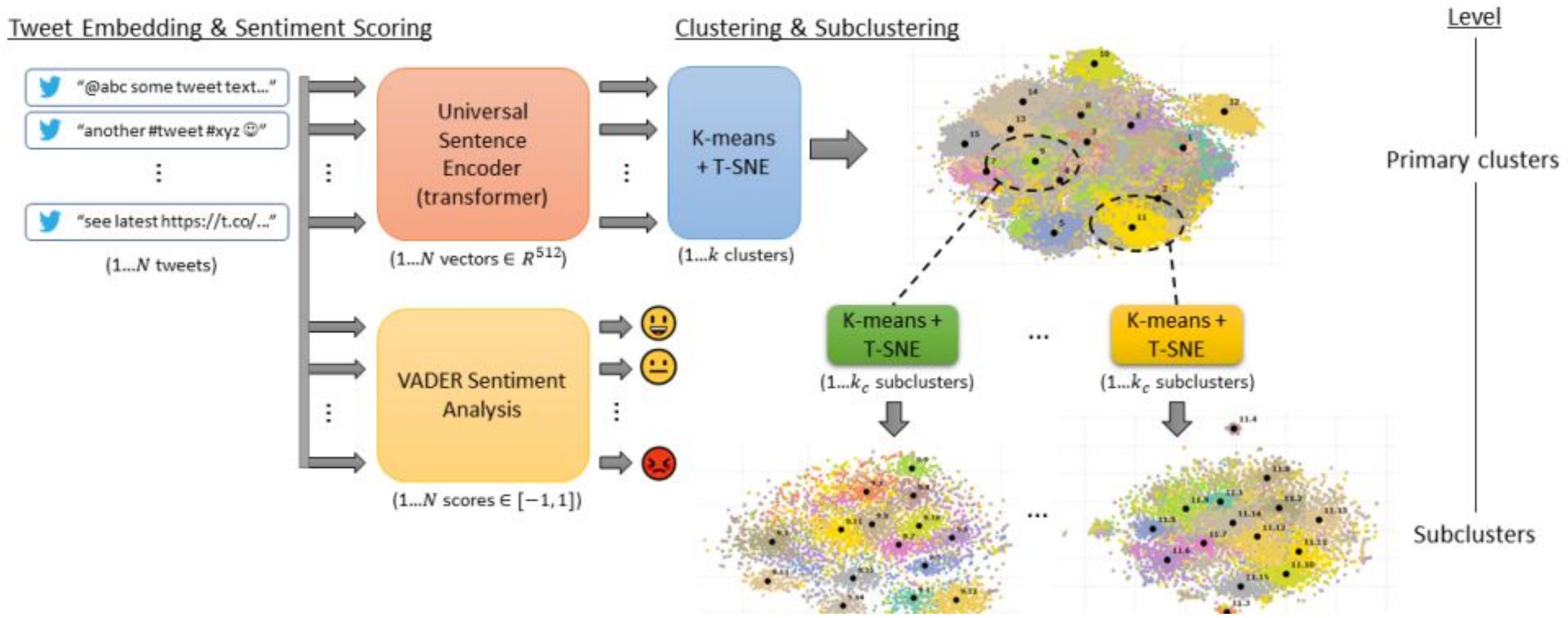
[2] Larson, Stefan & Mahendran, Anish & Lee, Andrew & Kummerfeld, Jonathan & Hill, Parker & Laurenzano, Michael & Hauswald, Johann & Tang, Lingjia & Mars, Jason. (2019). Outlier Detection for Improved Data Quality and Diversity in Dialog Systems. 517-527. 10.18653/v1/N19-1051.

[3] Qureshi, S., Hasanuzzaman, M., Saha, S., & Dias, G. (2019). The Verbal and Non Verbal Signals of Depression - Combining Acoustics, Text and Visuals for Estimating Depression Level. ArXiv, abs/1904.07656.

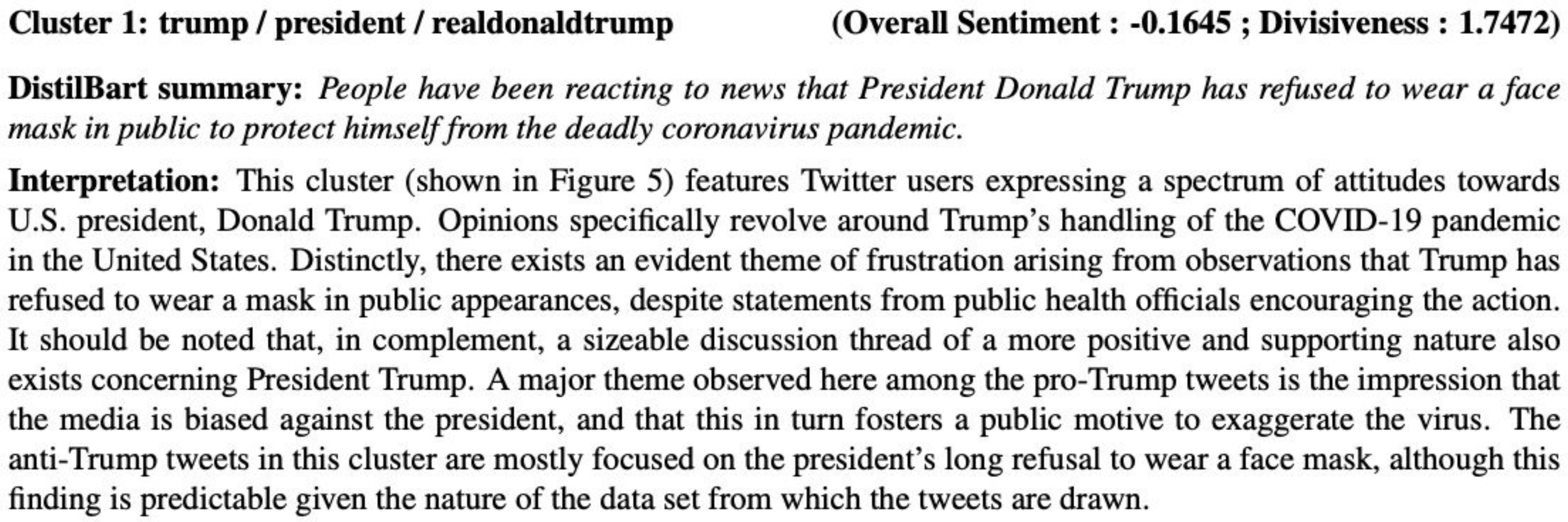
[4] Mihaylova, Tsvetomila & Karadzhov, Georgi & Atanasova, Pepa & Baly, Ramy & Mohtarami, Mitra & Nakov, Preslav. (2019). SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums.

Universal Sentence encoder - Use Cases

Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse



Pipeline



Cluster 1: trump / president / realdonaldtrump **(Overall Sentiment : -0.1645 ; Divisiveness : 1.7472)**

DistilBart summary: *People have been reacting to news that President Donald Trump has refused to wear a face mask in public to protect himself from the deadly coronavirus pandemic.*

Interpretation: This cluster (shown in Figure 5) features Twitter users expressing a spectrum of attitudes towards U.S. president, Donald Trump. Opinions specifically revolve around Trump's handling of the COVID-19 pandemic in the United States. Distinctly, there exists an evident theme of frustration arising from observations that Trump has refused to wear a mask in public appearances, despite statements from public health officials encouraging the action. It should be noted that, in complement, a sizeable discussion thread of a more positive and supporting nature also exists concerning President Trump. A major theme observed here among the pro-Trump tweets is the impression that the media is biased against the president, and that this in turn fosters a public motive to exaggerate the virus. The anti-Trump tweets in this cluster are mostly focused on the president's long refusal to wear a face mask, although this finding is predictable given the nature of the data set from which the tweets are drawn.

Results

Sanders, Abraham, Rachael White, Lauren Severson, Rufeng Ma, Richard McQueen, Haniel Campos Alcanatara Paulo, Yucheng Zhang, John S Erickson, und Kristin P Bennett. „Unmasking the Conversation on Masks: Natural Language Processing for Topical Sentiment Analysis of COVID-19 Twitter Discourse". Preprint. Health Informatics, 1. September 2020. https://doi.org/10.1101/2020.08.28.20183863.

Universal Sentence encoder - Limitations

- YouTube comments are noisy

- Discrepancy between data used for pretraining and our data

- Model has no understanding of the real world

Assignments for next week

1 Reading assignment

- Read Paper:
  - Sanders, Abraham, Rachael White, Lauren Severson, Rufeng Ma, Richard McQueen, Haniel Campos Alcanatara Paulo, Yucheng Zhang, John S Erickson, und Kristin P Bennett. „Unmasking the Conversation on Masks: Natural Language Processing for Topical Sentiment Analysis of COVID-19 Twitter Discourse“. Preprint. Health Informatics, 1. September 2020. https://doi.org/10.1101/2020.08.28.20183863.
- Answer the following questions in a summary of 150 words:
  - In which ‘mode of inquiry’ (Marres and Moats 2015) is the research project and paper operating?
  - What are the issues that are silenced / cannot be grasped by this mode of inquiry?

2 Language Model

- Download and setup the Jupyter notebook as described in our GitHub repository (https://github.com/FUB-HCC/seminar_critical-social-media-analysis)
- Preprocess and Embed your data with the Pipeline
- Pick some comments you found interesting in your prior analysis. Get similar comments.
- Answer the following questions in a summary of 150 words:
  - What can the model do well? When does it fail?
  - How can you use it in your project?
- Commit your Notebook with outputs to GitHub: create a new folder named [name]_assignment_session5
- Share your notebook URL in your assignment submission

Submit on Github (reply to issue) until 2 Dec 12h00 (noon)

What’s up next session?

Clustering and visualization!

Recommended readings

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

Cer, D.M., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder. ArXiv, abs/1803.11175.

open space

Feel free to approach us in case of questions…
(microphone or chat)