Text Mining

Core Methods in Educational Data Mining

Valdemar Švábenský | University of Pennsylvania | Nov 3, 2022

Based on the materials created by Ryan Baker for his EDM MOOC/MOOT

Previous assignment (Basic: PFA)

- Questions? Comments? Concerns?

Text mining

- **Automated extraction of information from text data**
- Related disciplines: natural language processing (NLP), discourse processing, computational linguistics…

- Difficult problem
- **Different than mining interaction/course data**, e.g.:
  - BKT/IRT work great for interaction data but less in text mining
  - SVM works great in text mining but less for interaction data

Characteristics of text data

- **Really high dimensionality**
  - Many words in a text corpus
- **Various levels of analysis** are possible
  - Individual phonemes/graphemes
  - Individual words (unordered or ordered)
  - Pairs (*bigrams*) or triplets (*trigrams*) of neighboring words
  - Sentences/paragraphs
  - Entire essays/books
- **Can you think of more characteristics?**

Applications of text mining in education

- Analysis of sentiment and emotions within learner utterances
  - (D'Mello et al., 2008)
- Studying content of online discussion forums
  - (Almatrafi et al., 2018)
- Studying pair collaboration online
  - (Dyke et al., 2013)
- Enhancing dialogues between students and tutoring systems
  - (Forsyth et al., 2013)
- **Can you think of more ways it could be used in education?**

Tools

- Python NLTK module (Natural Language Toolkit)
  - https://www.nltk.org/
- RapidMiner with its Text Processing extension
  - https://marketplace.rapidminer.com/UpdateServer/faces/product_details.xhtml?productId=rmx_text
- LightSide
  - https://www.cs.cmu.edu/~cprose/LightSIDE.html
  - Enables turning utterances into uni/bi/trigrams, as well as more powerful feature extraction, and then running ML on the data

Today’s topics

- BOW and TF–IDF
- LSA
- Semantic tagging
- Deep learning models
- Linguistic analysis

BOW and TF–IDF

Bag of words (BOW)

- **What is it?**
- **How would it look like on this dataset?**

  John likes to analyze data. Mary likes data analysis too.

- **When can it be useful in education?**

BOW example

(Input text)	how	are	you	do	thank
How are you?	1	1	1	0	0
How do you do?	1	0	1	2	0
Thank you.	0	0	1	0	1

Term frequency – Inverse document freq. (TF–IDF)

- **What is it?**
- **How would it look like on this dataset?**

  (Document 1) It will rain today.

  (Document 2) Today I will stay home.

- **When can it be useful in education?**

https://medium.com/analytics-vidhya/tf-idf-term-frequency-technique-easiest-explanation-for-text-classification-in-nlp-with-code-8ca3912e58c3

BOW and TF–IDF: discussion

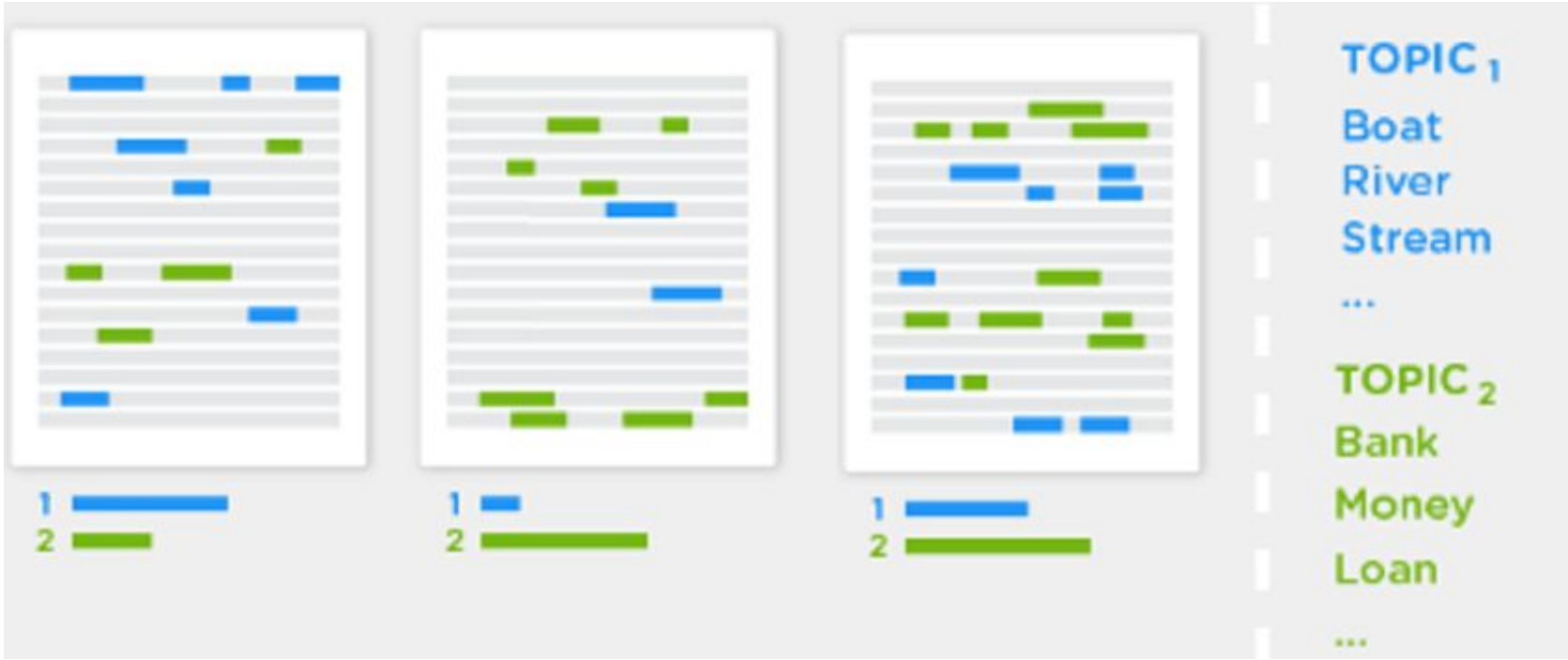
- What are the **advantages** of these approaches?
- What are the **disadvantages** of these approaches?

LSA

Latent semantic analysis (LSA)

- Also called *Latent semantic indexing (LSI)*
- Goal: find the **hidden topics** represented in documents
  - “How are the words related?”
  - Video illustration: https://upload.wikimedia.org/wikipedia/commons/transcoded/7/70/Topic_model_scheme.webm/Topic_model_scheme.webm.480p.vp9.webm

LSA: principle



Source: https://medium.com/swlh/spam-filtering-using-bag-of-words-aac778e1ee0b

LSA: example

Document	Cleaned document	Topic 1	Topic 2
He is a good dog.	good dog	0.3413	0.7199
The dog is too lazy.	lazy dog	0.3713	0.7089
That is a brown cat	brown cat	0.8609	-0.3659
The cat is active.	cat active	0.5166	-0.3850

Source: https://medium.com/swlh/spam-filtering-using-bag-of-words-aac778e1ee0b

LSA: data representation

- Sparse **“document-term” matrix**:
  - Each **row** is an **utterance** (a few words, a sentence, a paragraph)
  - Each **column** is a **word** that can be present (1) or absent (0)
- **Does not model syntax**, just word presence (like BOW)
  - (Landauer, Foltz, & Laham, 1998)

Example 1: https://towardsdatascience.com/latent-semantic-analysis-intuition-math-implementation-a194aff870f8
Example 2: https://www.datacamp.com/tutorial/discovering-hidden-topics-python

LSA: implementation

- Conducts **singular value decomposition** of the document-term matrix
  - Matrix factorization technique
  - Conceptually similar to factor analysis
- Goal: identify patterns in the **relationships between the terms and latent concepts**
- **Is it supervised or unsupervised?**

Math explained: https://matpalm.com/lsa_via_svd/eg1.html

LSA: discussion

Discuss in small groups (2–3 people):

- What are the **educational applications** of LSA?
- What are the **advantages** of LSA?
- What are the **disadvantages** of LSA?

Semantic tagging

Semantic tagging

- Reduces words to **semantic categories**
  - E.g., “negative emotion” ← hurt, scared, sad, …
- Analysis is then **less dependent on specific words**
- Two popular taggers (software tools):
  - **LIWC** (Linguistic Inquiry and Word Count): https://www.liwc.app/
  - **Wmatrix**: https://ucrel.lancs.ac.uk/wmatrix/

Semantic tagging: LIWC example (input)

It is a period of civil war. Rebel spaceships, striking from a hidden base, have won their first victory against the evil Galactic Empire.

During the battle, Rebel spies managed to steal secret plans to the Empire's ultimate weapon, the DEATH STAR, an armored space station with enough power to destroy an entire planet.

Pursued by the Empire's sinister agents, Princess Leia races home aboard her starship, custodian of the stolen plans that can save her people and restore freedom to the galaxy....

Source: https://imsdb.com/scripts/Star-Wars-A-New-Hope.html

Semantic tagging: LIWC example (output)

Traditional LIWC Dimension	Your Text	Average for Story Language
I-words (I, me, my)	0.00	3.22
Positive Tone	3.57	2.18
Negative Tone	7.14	1.75
Social Words	5.95	10.50
Cognitive Processes	5.95	8.70

Semantic tagging: discussion

Discuss in small groups (different than before):

- What are the **educational applications**?
- What are the **advantages**?
- What are the **disadvantages**?
- When is semantic tagging better than looking for specific words? When is it worse?

Deep learning models

Deep learning

- Complex neural networks
  - We focus on transformer (foundation) models – good for sequential data (text)
- + Can be very accurate
- – Blackbox: sacrifices model explainability

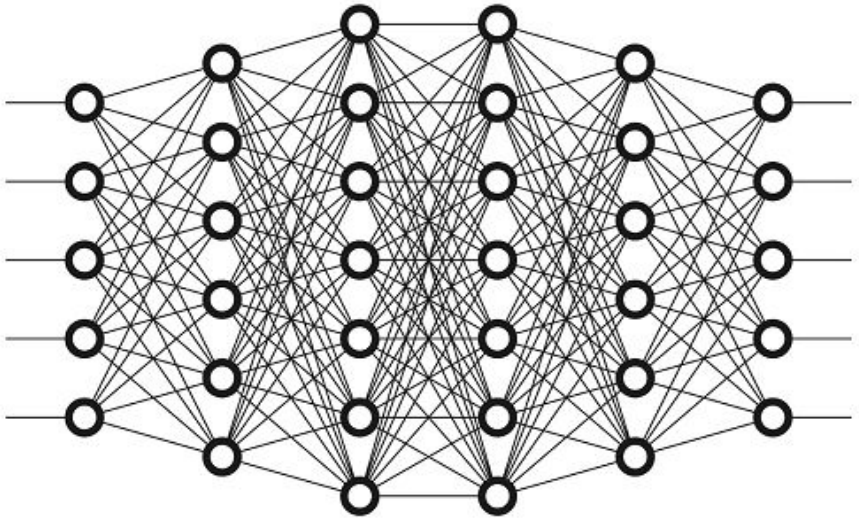


Image source: https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063

Embedding

- Representation of any text (word, sentence) as a feature vector of a fixed dimension
- **Why can this be useful?**

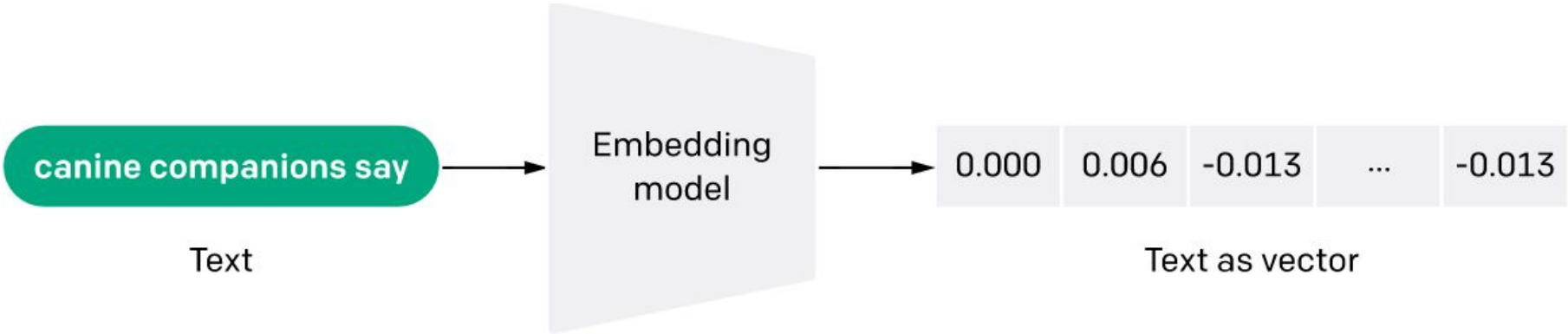


Image source: https://openai.com/blog/introducing-text-and-code-embeddings/

Usage example: sentence similarity

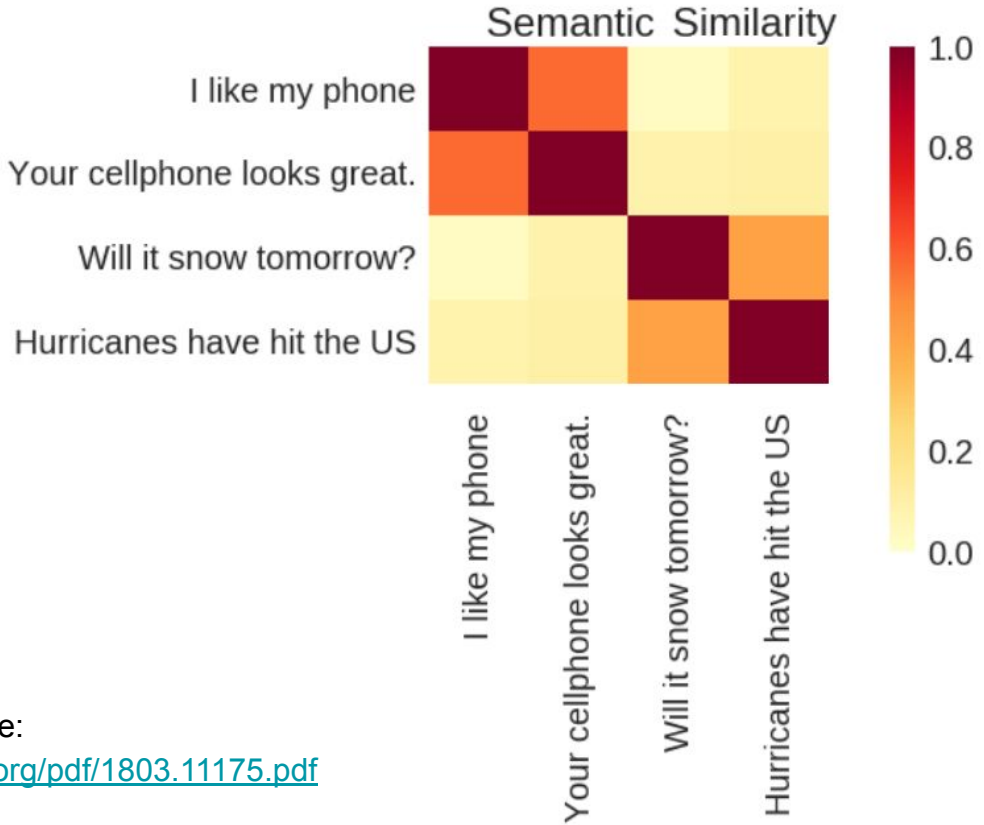


Image source:
https://arxiv.org/pdf/1803.11175.pdf

Universal sentence encoder (USE)

- Google, since 2018
  - https://arxiv.org/pdf/1803.11175.pdf
- 512-dimensional embeddings
- Trained on various texts (unspecified which)
- Full models available for download (now v4/5)
  - https://tfhub.dev/google/universal-sentence-encoder-large/5
- **Can you think of any educational applications?**

Sentence BERT (SBERT)

- Darmstadt University, since 2019
  - https://arxiv.org/pdf/1908.10084.pdf
- 768-dimensional embeddings
- Trained on books and Wikipedia
- Full models available for download (several languages)
  - https://www.sbert.net
  - Usable as-is in Python (few lines of code) or modifiable
  - HuggingFace (https://huggingface.co/) – wrapper around pre-built/pre-trained models, including for SBERT

Generative Pre-trained Transformer 3 (GPT-3)

- Open AI, since 2020
  - https://arxiv.org/pdf/2005.14165.pdf
- 768 (and larger)-dimensional embeddings
- Trained on existing text datasets, books and Wikipedia
- Usable in Python (account needed)
  - https://openai.com/api/
- **Usage**: predicting the next word, generating new text

Word embedding vs. sentence-level embedding

- Context-free models (like **word2vec**) generate a single embedding for each word
  - The word “right” would have the same representation in “I’ll make the payment right away” and “Take a right turn”
- **USE** and **BERT** operate on the sentence level, generating embeddings based on the context
  - https://eng.zemosolabs.com/text-classification-bert-vs-dnn-b226497c9de7

Linguistic analysis

TAALES, TAACO

- Tools for automated analysis of
  - Lexical sophistication (e.g., age of exposure)
  - Cohesion
  - … and much more
  - https://www.linguisticanalysistools.org/tools.html
- **Can you think of any educational applications?**

Text coherence

- “How hard is a text to read?”
- Newer version of reading-level metrics
  - E.g., Fleisch-Kincaid
- Coh-Metrix
  - Tool that provides many metrics about a text, incl. coherence
  - http://cohmetrix.com/
- Requires several grammatically correct sentences (e.g., in essays), not suitable for short pieces of text

Quiz time!

- On your phone, go to **play.blooket.com**
- Enter the ID code shown on the projector
- Choose your nickname (SFW please) and avatar
- Answer multiple-choice questions: both accuracy and speed count

Image credit: https://www.blooket.com/