TEXT MINING

L08. TOPIC MODELLING & SUMMARIZATION

SUZAN VERBERNE 2022

Universiteit Leiden

QUIZ ABOUT WEEK 7

- Name the kind of task that is used to learn word embeddings (e.g. by pre-training a BERT model)
    a. Sequential processing
    b. Language modelling
    c. Recurrence
    d. Semantic role labelling

QUIZ ABOUT WEEK 7

- Name the kind of task that is used to learn word embeddings (e.g. by pre-training a BERT model)
  a. Sequential processing
  b. Language modelling
  c. Recurrence
  d. Semantic role labelling

QUIZ ABOUT WEEK 7

- Which statements about context in sequential mdoels are true?
  a. LSTMs are sequential models with longer memories than traditional RNNs
  b. BERT models can make use of longer contexts than LSTMs
  c. The attention mechanism in Transformer models has quadratic complexity relative to the input length
  d. The maximum input length for BERT models is limited by computational memory

QUIZ ABOUT WEEK 7

- Which statements about context in sequential mdoels are true?
  a. LSTMs are sequential models with longer memories than traditional RNNs
  b. BERT models can make use of longer contexts than LSTMs
  c. The attention mechanism in Transformer models has quadratic complexity relative to the input length
  d. The maximum input length for BERT models is limited by computational memory

QUIZ ABOUT WEEK 7

- Consider a named entity recognition task for Chinese texts with limited labeled data (300 documents, 3000 entities). What do we need to build a classifier using transfer learning?

  a. A pre-trained Chinese BERT model

  b. A large collection of Chinese texts to train a BERT model

  c. GPU computing, a lot of time, and the 3000 items for supervised learning

  d. GPU computing, a pre-trained Chinese BERT model and the 3000 items for supervised fine-tuning

QUIZ ABOUT WEEK 7

- Consider a named entity recognition task for Chinese texts with limited labeled data (300 documents, 3000 entities). What do we need to build a classifier using transfer learning?

  a. A pre-trained Chinese BERT model
  b. A large collection of Chinese texts to train a BERT model
  c. GPU computing, a lot of time, and the 3000 items for supervised learning
  d. GPU computing, a pre-trained Chinese BERT model and the 3000 items for supervised fine-tuning

TODAY'S LECTURE

- Topic modelling
  - LDA
- Automatic summarization
  - Extractive summarization methods
  - Abstractive summarization methods
  - Challenges
  - Evaluation

TOPIC MODELLING

(BRIEFLY MENTIONED IN J&M CHAPTER 6)

TOPIC MODELLING

- Assumptions
  - each document consists of a mixture of *topics*, and
  - each *topic* consists of a mixture of words

- "Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives" (Blei, 2009)

- Often used for data exploration: what topics are discussed and how frequent are they?

RAW EXAMPLE OUTPUT

```
Topic #1: oxygen urban city burning gaseous phi wetting films ethylene cities
Topic #2: hydrogen pvt thermodynamically splitting cryogenic losses recorded convert upgrading maxima
Topic #3: pure calibration furnace machining plots france inflation shapes plot coke
Topic #4: pollution south bay meso-beta stainless idealized bengal atlantic out-of-plane slag
Topic #5: dispersion turbulent one-dimensional bundles islands preform accepted character two-dimensional interphase
Topic #6: rankine organic season seasonal pcm dry plume abundance seasons gas
Topic #7: instability code weak true mmp extra displacements codes trigger mese-nh
Topic #8: calculation dust eastern elevated heights middle clc transferred rank adjacent
Topic #9: parametric kinetic forming geometry degrees operated enhancement forced characterised molten
Topic #10: carbonation looping tropospheric gev entrainment fluidization counter supplies alpha resolve
Topic #11: solid defect bottom transportation supply mixed exact reductions evaporators facility
Topic #12: gas energy micro-scale macro-scale system production geothermal efficiency material materials
Topic #13: vortex mixtures multi-objective refrigerants cutting optimum preferred inverse recuperator feedstock
Topic #14: mixture crack beams vibration beam rotation moduli meso-structures tensile recycling
Topic #15: velocity assimilation rock rocks swirl sound short-term load-displacement box comparatively
Topic #16: offshore impure corrosion gauge written constructal nickel subroutines discontinuous inclination
Topic #17: ground cascade anthropogenic quantified logic cycle planar record usefulness graphical
Topic #18: methanol sectors economy alternatives thrust neglected august planetary sector southwest
Topic #19: arc oil saturated viscous concentrations constituents boiling emulsion apoa-iv resistances
Topic #20: discharge scs china pem soil accomplished iron subsurface yield shore
Topic #21: sea soil ecosystem coastal synoptic ecological ecosystems variability nitrogen marine
Topic #22: electricity injection investment aim emissions price pressure warming hourly building
Topic #23: compression dissipation regenerator assemblages coefficients cyclones reanalysis core-periphery carriers a
Topic #24: meso preheating sludge mounted surroundings sdms distilled transported characterisation experiment
Topic #25: thermoelectric unit configurations temperatures mode hybrid heat subroutine lowest configuration
```

TOPIC MODELLING

- Topic modelling is an unsupervised technique
  - Topic labels are not given
  - The number of topics needs to be pre-specified
  - Think about it as clustering

LATENT DIRICHLET ALLOCATION (LDA)

- Most used topic modelling technique

[PDF] **Latent dirichlet allocation**

DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org

We describe **latent Dirichlet allocation** (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in …

☆ Save 99 Cite Cited by 44615 Related articles All 97 versions Web of Science: 17559 »

- Popular in applied Text Mining in many domains
  - (especially the social sciences)

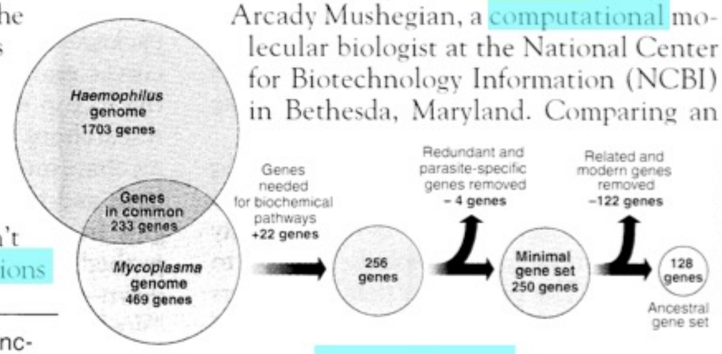
INTUITION BEHIND LDA

- One document exhibits multiple topics
- Words are more related to one topic than to another
  - Blue: Computer analysis
  - Pink: Evolutionary biology
  - Yellow: Genomics

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



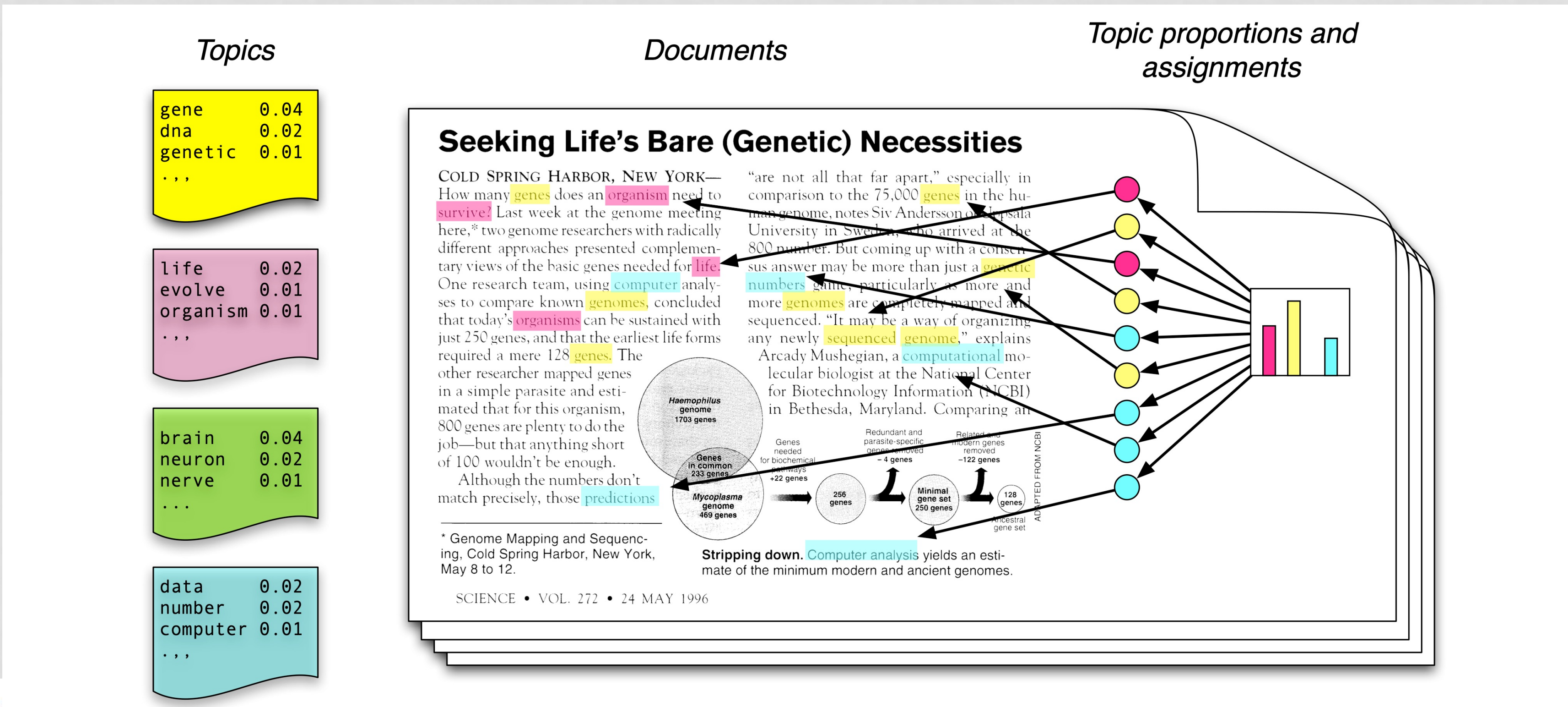
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

GENERATIVE MODEL

- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics



GENERATIVE PROBABILISTIC MODEL

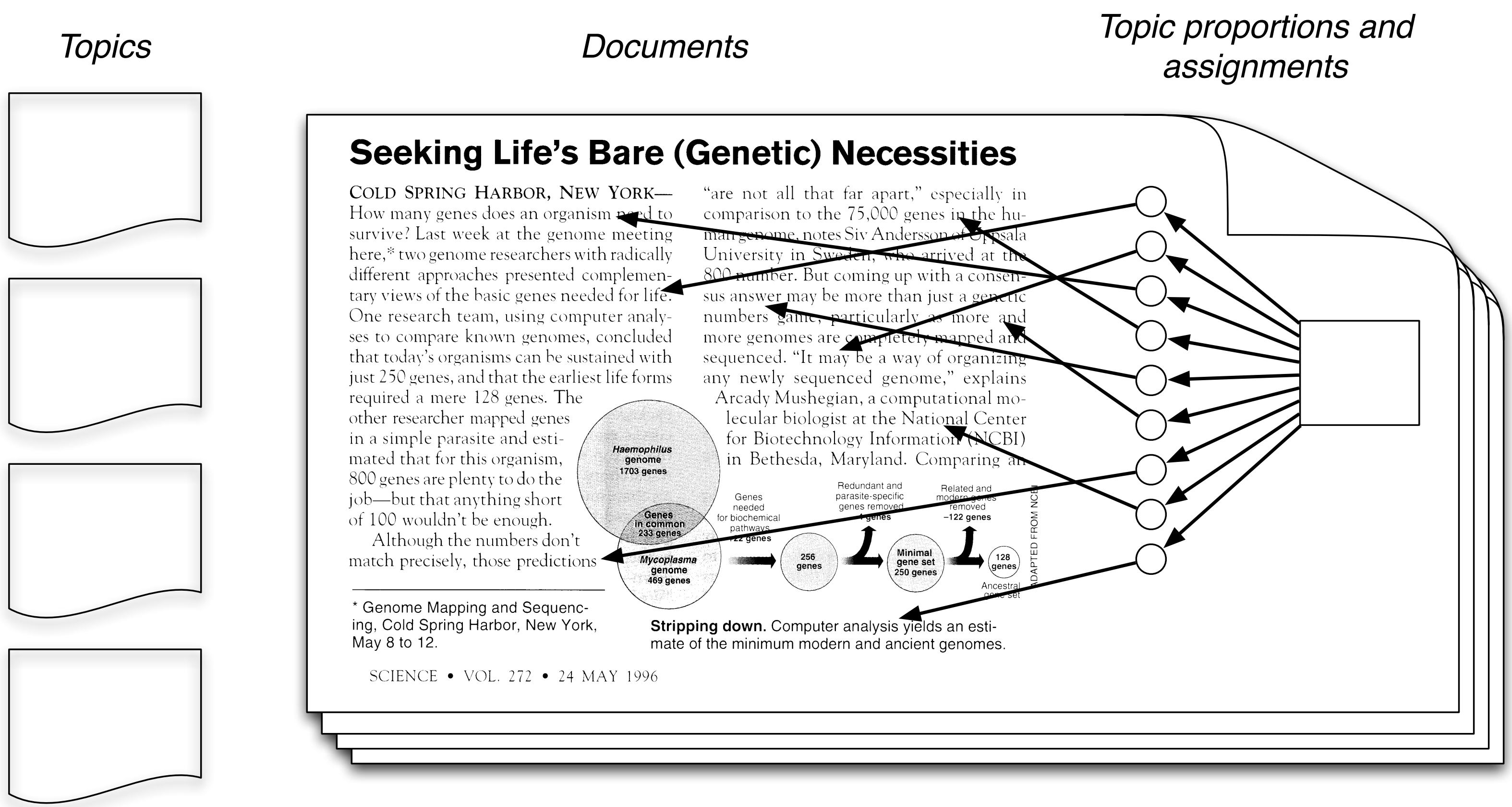
- Topic = probability distribution over fixed vocabulary. Every topic contains a probability for every word in the vocabulary
- Each document is a distribution over topics
  - Dirichlet distribution: continuous multivariate probability distribution
  - The prior set on the Dirichlet distribution is sparse
    - Assumption: each document covers only a small set of topics and each topic uses only a small set of words frequently

GENERATIVE MODEL

- We generate a document as a bag of words:
  - We draw a topic from the distribitution (e.g. yellow), lookup the yellow distribution, and draw a word from the yellow distribution. Etc.
  - The order of words doesn't matter; the words are drawn independently of each other

- We only observe the words in the documents. The topics are latent.

The posterior distribution



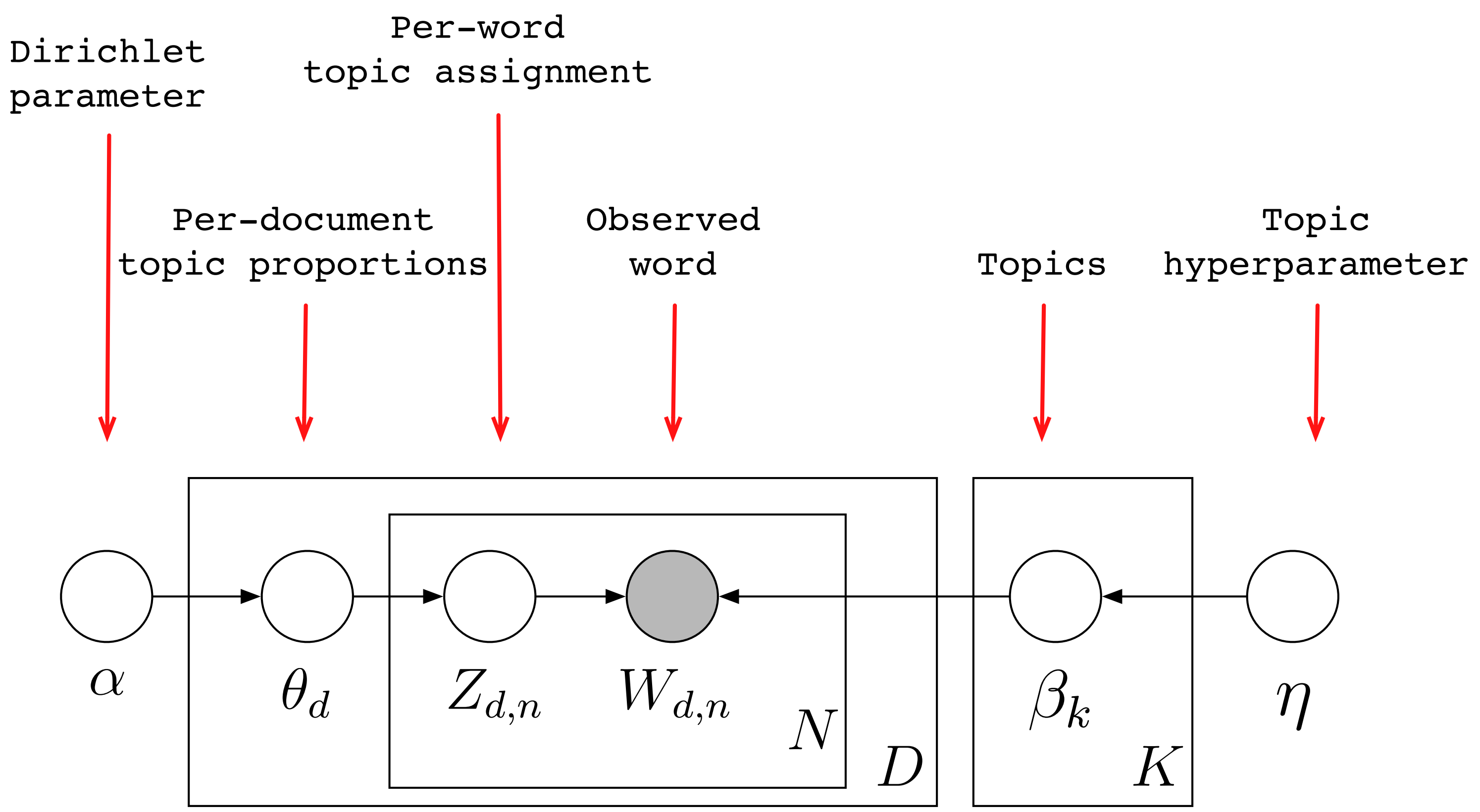
- In reality, we only observe the documents
- Our goal is to **infer** the underlying topic structure

LEARN THE DISTRIBUTIONS

1. What are the topics, what are the distributions over terms?
2. For each document, what is the distribution over topics?

Universiteit Leiden

Latent Dirichlet allocation



LEARN THE TOPICS FROM THE DATA

- Goal: to learn β (the topic models) and θ
- We only observe the words W
- Start with random probability distributions of words in topics and of topics in documents
- Update the probability distributions while observing the words in the documents (Bayesian framework)
- Until β converges, or the maximum number of epochs has been reached

Universiteit Leiden

Good explanation of the training process:
https://highdemandskills.com/topic-modeling-intuitive/

LATENT DIRICHLET ALLOCATION (LDA)

- Popular implementation in gensim
- https://radimrehurek.com/gensim/models/ldamodel.html

```
# extract 100 LDA topics, updating once every 10,000
lda = LdaModel(corpus=mm, id2word=id2word, num_topics=100,
update_every=1, chunksize=10000, passes=1)

# use LDA model: transform new doc to bag-of-words, then apply lda
doc_bow = doc2bow(document.split())
doc_lda = lda[doc_bow]

# doc_lda is vector of length num_topics representing weighted
presence of each topic in the doc
```

RAW EXAMPLE OUTPUT

```
Topic #1: oxygen urban city burning gaseous phi wetting films ethylene cities
Topic #2: hydrogen pvt thermodynamically splitting cryogenic losses recorded convert upgrading maxima
Topic #3: pure calibration furnace machining plots france inflation shapes plot coke
Topic #4: pollution south bay meso-beta stainless idealized bengal atlantic out-of-plane slag
Topic #5: dispersion turbulent one-dimensional bundles islands preform accepted character two-dimensional interphase
Topic #6: rankine organic season seasonal pcm dry plume abundance seasons gas
Topic #7: instability code weak true mmp extra displacements codes trigger mese-nh
Topic #8: calculation dust eastern elevated heights middle clc transferred rank adjacent
Topic #9: parametric kinetic forming geometry degrees operated enhancement forced characterised molten
Topic #10: carbonation looping tropospheric gev entrainment fluidization counter supplies alpha resolve
Topic #11: solid defect bottom transportation supply mixed exact reductions evaporators facility
Topic #12: gas energy micro-scale macro-scale system production geothermal efficiency material materials
Topic #13: vortex mixtures multi-objective refrigerants cutting optimum preferred inverse recuperator feedstock
Topic #14: mixture crack beams vibration beam rotation moduli meso-structures tensile recycling
Topic #15: velocity assimilation rock rocks swirl sound short-term load-displacement box comparatively
Topic #16: offshore impure corrosion gauge written constructal nickel subroutines discontinuous inclination
Topic #17: ground cascade anthropogenic quantified logic cycle planar record usefulness graphical
Topic #18: methanol sectors economy alternatives thrust neglected august planetary sector southwest
Topic #19: arc oil saturated viscous concentrations constituents boiling emulsion apoa-iv resistances
Topic #20: discharge scs china pem soil accomplished iron subsurface yield shore
Topic #21: sea soil ecosystem coastal synoptic ecological ecosystems variability nitrogen marine
Topic #22: electricity injection investment aim emissions price pressure warming hourly building
Topic #23: compression dissipation regenerator assemblages coefficients cyclones reanalysis core-periphery carriers
Topic #24: meso preheating sludge mounted surroundings sdms distilled transported characterisation experiment
Topic #25: thermoelectric unit configurations temperatures mode hybrid heat subroutine lowest configuration
```

CHALLENGES WITH LDA

- Choose the number of topics
- Random initiation of the clustering → the outcome is non-deterministic
  - Alternative model: Non-negative matrix factorization (NMF)
- Interpreting the output: what do the topics mean?

EXERCISE

1. Topic modelling is often used as method for data exploration (followed by expert interpretation).

   How could you use the output of LDA for text classification instead? Think about the case of classifying health-related newspaper articles.

2. How would you evaluate an LDA model?

Discuss with your neighbour(s)

EVALUATION OF TOPIC MODELLING

- Use LDA output for classification?
  - Represent document as 'bag of topics' (a vector with the topic ids as features and the topic probability as value)
- How to evaluate?
  - Topic coherence, also used for optimizing the number of topics
    - E.g. use a word2vec model to measure similarity of words inside a topic and between topics
  - Human evaluation:
    - Word intrusion: given these 5 high-probability topic words + 1 random word, can you find the intruder?

Universiteit Leiden

O'callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*(13), 5645-5657.

MORE INFORMATION

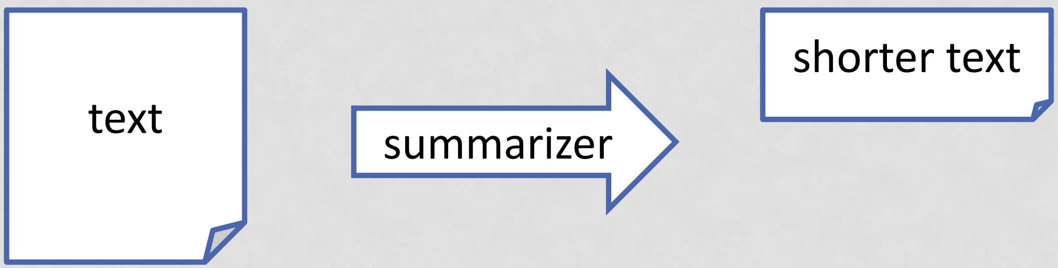
- Additional resources for who is interested in topic modelling
  - Tutorial (optional): https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
  - LDA lecture: http://videolectures.net/mlss09uk_blei_tm/
  - Alternative to LDA: Non-negative matrix factorization (NMF) https://towardsdatascience.com/topic-modeling-articles-with-nmf-8c6b2a227a45

AUTOMATIC SUMMARIZATION

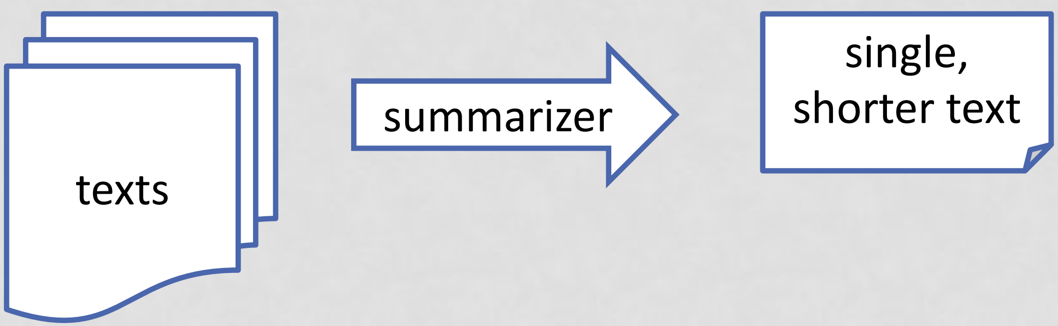
Universiteit Leiden

SINGLE-DOCUMENT VS. MULTI-DOCUMENT

- Single-document summarization



- Multi-document summarization



SINGLE-DOCUMENT VS. MULTI-DOCUMENT

- Single-document summarization:
  - News articles
  - Scientific articles
  - Meeting reports (minutes)
- Multi-document summarization:
  - Output of a search engine
  - News about a single topic from multiple sources
  - Discussion threads summarization (“most commenters say that…”)

EXTRACT VS. ABSTRACT

- An extract is a summary composed completely of material from the source
- An abstract is a summary that contains material not originally in the source, but shorter paraphrases

SUMMARIZATION METHODS

EXTRACTIVE VS ABSTRACTIVE

Extractive summarization

Select the most important nuggets (sentences)

This is a classification or ranking task

- Classification: for each sentence, select it: yes/no
- Ranking: assign a score to each sentence, then select the top-k

EXTRACTIVE VS ABSTRACTIVE

Extractive summarization

Select the most important nuggets (sentences)

This is a classification or ranking task

- Classification: for each sentence, select it: yes/no
- Ranking: assign a score to each sentence, then select the top-k

Abstractive summarization

Learn a text-to-text-transformation model (cf. translation)

- Training data: pairs of longer and shorter texts
- Sequence-to-sequence models: Learning a mapping between an input sequence and an output sequence

EXTRACTIVE VS ABSTRACTIVE

Extractive summarization

- Feasible / easy to implement
- Reliable (literal re-use of text)
- But limited in terms of fluency
- (Fixes required after sentence selection)

EXTRACTIVE VS ABSTRACTIVE

Extractive summarization

- Feasible / easy to implement
- Reliable (literal re-use of text)
- But limited in terms of fluency
- (Fixes required after sentence selection)

Abstractive summarization

- More natural/fluent result
- But a lot of training data needed
- And risk of untrue content

BASELINE SUMMARIZATION SYSTEM

- Take the first three sentences from the document
  - Strong baseline!
  - “State-of-the- art models only slightly outperform the Lead-3 baseline, which generates summaries by extracting the first three sentences of the source document.” (Kryściński et al, 2019)

EXTRACTIVE SUMMARIZATION

SENTENCE SELECTION METHODS

- Unsupervised methods:
  - Centrality-based
  - (Graph-based)
- Supervised methods:
  - Feature-based
  - (Embeddings based)

UNSUPERVISED SENTENCE SELECTION

- Centrality-based methods for sentence selection
  - Measure the cosine similarity between each sentence and the document
    - Either using the sparse vector space with words as dimensions
    - Or the dense vector space using embeddings representations
  - Select the sentences with the highest similarity (the most representative sentences)

SUPERVISED SENTENCE SELECTION

- Feature engineering + classifier (e.g. SVM)
- Features:
  - position in the document
  - word count
  - word lengths
  - word frequencies
  - punctuation
  - representativeness (similarity to full document/title)
  - etc.

PROBLEMS WITH SENTENCE SELECTION

- Selecting sentences that contain unresolved references to sentences not included in the summary or not explicitly included in the original document:
  - “Our investigations have shown this to be true.”
  - “There are three distinct methods to be considered.”

PROBLEMS WITH SENTENCE SELECTION

- Selecting sentences that contain unresolved references to sentences not included in the summary or not explicitly included in the original document:
  - “Our investigations have shown this to be true.”
  - “There are three distinct methods to be considered.”

- Improvements that might be needed after sentence selection:
  - Sentence ordering
  - Sentence revision
  - Sentence fusion
  - Sentence compression

ABSTRACTIVE SUMMARIZATION

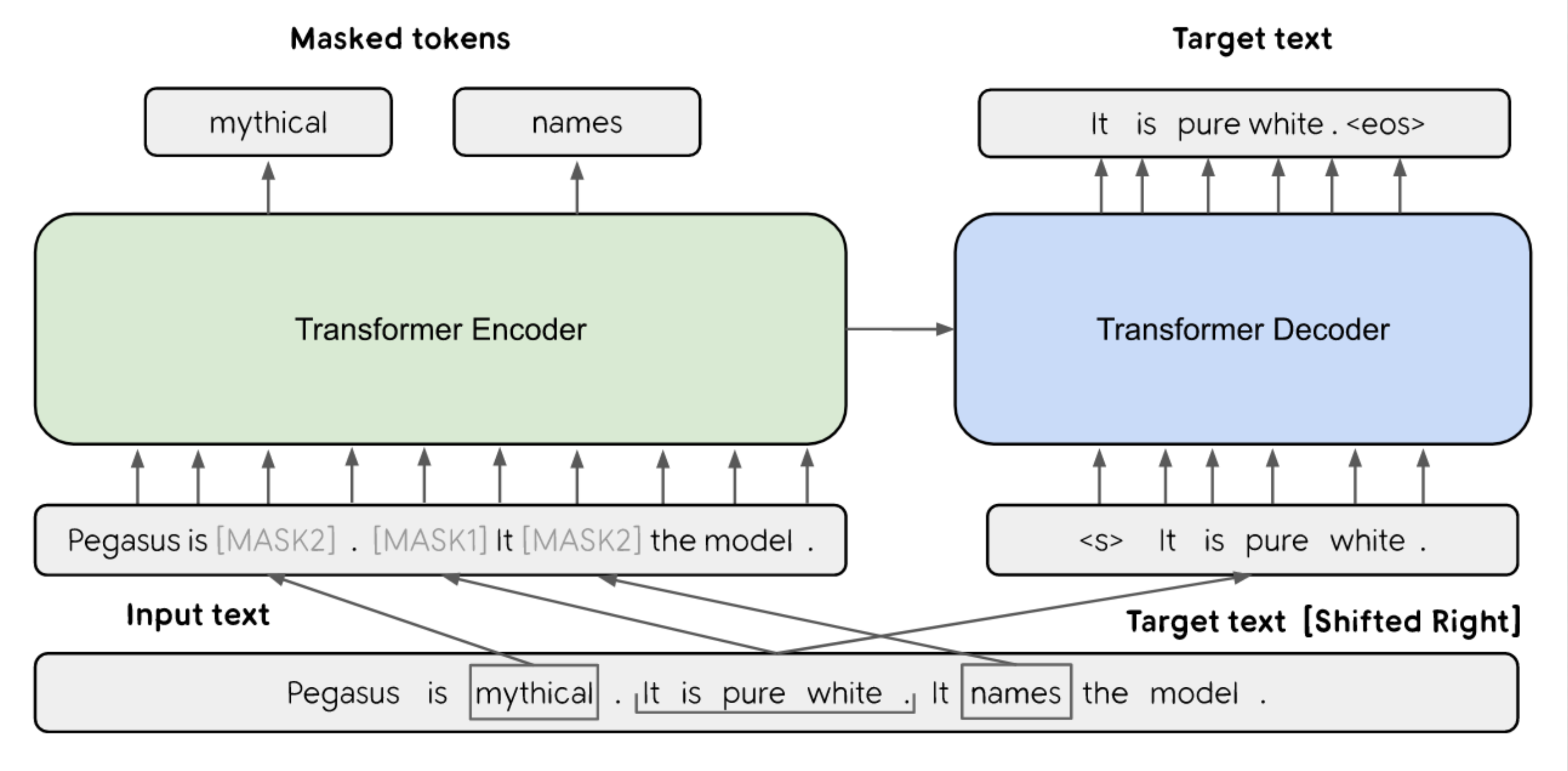
SEQUENCE-TO-SEQUENCE LEARNING

Summarization as translation task:

- Translate a longer text into a shorter text
- Sequence-to-sequence models; encoder-decoder architectures (RNNs/LSTMs/Transformers)
- Training data: pairs of longer and shorter texts. Examples:
  - Professionally written summaries for benchmarking purposes
  - For scientific documents: full article and abstract
  - Editor-written summaries of comment threads (NY Times)
  - User-generated content: Reddit TL;DR summaries

ENCODER-DECODER PRE-TRAINING

- PEGASUS: encoder-decoder pre-training for summarization



Zhang et al. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning* (pp. 11328-11339)

PEGASUS

- Pre-training objectives (self-supervised):
  1. Masked Language Modelling (like BERT)
  2. Gap Sentences Generation (GSG)

- Motivation:
  - Large-scale document-summary datasets (for supervised learning) are rare;
  - Creating training data is expensive (‘low-resource summarization’)

- Pre-training with gap sentences followed by fine-tuning with small numbers of supervised pairs, gives “state-of-the-art results in 6 datasets with only 1000 examples”

CHALLENGES OF SUMMARIZATION

RESEARCH CHALLENGES

- Factual consistency (for abstractive summarization)
- Task subjectivity/ambiguity
- Training data bias
- Evaluation

Universiteit Leiden

FACTUAL CONSISTENCY

- Factual consistency is the main challenge of abstractive summarization models
  - “appearance of hallucinations in generated abstracts”

FACTUAL CONSISTENCY

Original	Generated summary
a man flipping in the air with a snowboard above a snow covered hill	A snowboarder is doing a trick on a snowy slope .
many toilets without its upper top part near each other on a dark background	A row of toilets sitting on a tiled floor .
A table with three place settings with meat , vegetables and side dishes on it	A table topped with plates of food and a glass of wine .
A woman is leaning over a toilet , while her arms are inside a lawn and garden trash bag .	A woman is cleaning a toilet in a park .

FACTUAL CONSISTENCY

Article
Quick-thinking: Brady Olson, a teacher at North Thurston High, took down a gunman on Monday. A Washington High School teacher is being hailed a hero for tackling a 16-year-old student to the ground after he opened fire on Monday morning (...)
Summary - Factually incorrect
Brady Olson, a Washington High School teacher at North Thurston High, opened fire on Monday morning. No one was injured after the boy shot twice toward the ceiling in the school commons before classes began at North Thurston High School in Lacey (...)

Table 4: Example of a factually incorrect summary generated by an abstractive model. Top: ground-truth article. Bottom: summary generated by model.

Universiteit Leiden

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

FACTUAL CONSISTENCY

- Factual consistency is the main challenge of abstractive summarization models
  - “appearance of hallucinations in generated abstracts”
- A paper from 2020 showed that the majority of generated summaries contained non-faithful content
- Human judgement is still crucial for this kind of evaluation as automatic metrics do not strongly correlate with summary faithfulness

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020, July). On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906-1919).

TASK SUBJECTIVITY/AMBIGUITY

Article

The glowing blue letters that once lit the Bronx from above Yankee stadium failed to find a buyer at an auction at Sotheby's on Wednesday. While the 13 letters were expected to bring in anywhere from $300,000 to $600,000, the only person who raised a paddle - for $260,000 - was a Sotheby's employee trying to jump start the bidding. The current owner of the signage is Yankee hall-of-famer Reggie Jackson, who purchased the 10-feet-tall letters for an undisclosed amount after the stadium saw its final game in 2008. No love: 13 letters that hung over Yankee stadium were estimated to bring in anywhere from $300,000 to $600,000, but received no bids at a Sotheby's auction Wednesday. The 68-year-old Yankee said he wanted 'a new generation to own and enjoy this icon of the Yankees and of New York City.', The letters had beamed from atop Yankee stadium near grand concourse in the Bronx since 1976, the year before Jackson joined the team. (...)

Unconstrained Summary A

There was not a single buyer at the auction at Sotheby's on Wednesday for the glowing blue letters that once lit the Bronx's Yankee Stadium. Not a single non-employee raised their paddle to bid. Jackson, the owner of the letters, was surprised by the lack of results. The venue is also auctioning off other items like Mets memorabilia.

Unconstrained Summary B

The once iconic and attractive pack of 13 letters that was placed at the Yankee stadium in 1976 and later removed in 2008 was unexpectedly not favorably considered at the Sotheby's auction when the 68 year old owner of the letters attempted to transfer its ownership to a member the younger populace. Thus, when the minimum estimate of $300,000 was not met, a further attempt was made by a former player of the Yankees to personally visit the new owner as an

Universiteit Leiden

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

TRAINING DATA BIAS

- Most used benchmark sets for training and evaluating summarization models are based on news data
  - E.g. the CNN/DailyMail (Nallapati et al., 2016) dataset

TRAINING DATA BIAS

- Most used benchmark sets for training and evaluation summarization models are based on news data
  - E.g. the CNN/DailyMail (Nallapati et al., 2016) dataset
- In newspaper articles, the most important information is in the first paragraph
  - This is used as a feature in summarization models
  - And this is why Lead-3 is such a strong baseline
- In other domains than newspaper data (e.g. books, legal documents, reviews), this characteristic does not always apply

EVALUATION

HOW WOULD YOU EVALUATE A SUMMARIZER?

- You have developed a summarization model
- How do you evaluate it?

EVALUATION OF SUMMARIZATION

- Compare to reference summaries
- Ask human judges

COMPARE TO REFERENCE SUMMARIES

- Compute overlap with human reference summary
- ROUGE metrics
  - "Recall-Oriented Understudy for Gisting Evaluation"
  - Measures quality of a summary by comparison with reference summaries (literal)

COMPARE TO REFERENCE SUMMARIES

- ROUGE: the proportion of n-grams from the reference summaries that occur in the automatically created summary (‘recall-oriented’)

$$\text{ROUGE}-\text{N} = \frac{\#\ \text{n}-\text{grams in automatic AND reference summary}}{\#\ \text{n}-\text{grams in reference summary}}$$

  - ROUGE-1: overlap of unigrams (single words) between the system and reference summaries
  - ROUGE-2: overlap of bigrams (word pairs) between the system and reference summaries
  - ROUGE-L: overlap of Longest Common Subsequences (LCS); identifies longest n-grams that co-occur in both summaries

COMPARE TO REFERENCE SUMMARIES

Toy example:

- Reference Summary: police killed the gunman
- System A: police kill the gunman
- System B: the gunman kill police

ROUGE-2:

COMPARE TO REFERENCE SUMMARIES

Toy example:

- Reference Summary: police killed the gunman
- System A: police kill the gunman
- System B: the gunman kill police

ROUGE-2

- A: \$ police, police kill, kill the, the gunman, gunman \$ → 3/5
- B: \$ the, the gunman, gunman kill, kill police, police \$ → 1/5
- → System A is better than system B according to ROUGE-2

EVALUATION OF SUMMARIZATION

Method:

- Compare to reference summaries → ROUGE
- Ask human judges

ASK HUMAN JUDGES TO GRADE SUMMARIES

- Criteria to rate a summary:
  - Relevance/importance: selection of important content from the source
  - Consistency: factual alignment between the summary and the source
  - Fluency: quality of individual sentences
  - Coherence: collective quality of all sentences
- Ask multiple judges per summary

CHALLENGES IN EVALUATION (ABSTRACTIVE)

- ROUGE often has weak correlation with human judgments
- But human judgments for relevance (importance) and fluency are strongly correlated to each other

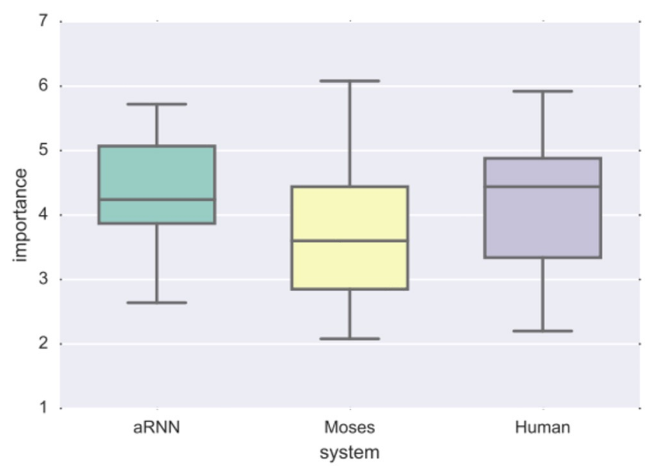


Figure 2: Importance scores given by human subjects to the wo systems and human description.

Figure 3: Fluency scores given by human subjects to the two ystems and human description.

Universiteit Leiden

Wubben et al. (2016). Abstractive Compression of Captions with Attentive Recurrent Neural Networks. Proceedings of the 9th International Natural Language Generation conference (INLG)

CONCLUSIONS

HOMEWORK

- Read:
  - Zhang et al. (2020) PEGASUS: pre-training with extracted gap-sentences for abstractive summarization
- Complete assignment 2: information extraction (CRF)
- Send in via Brightspace before or on Monday November 14:
  - Submit your report as PDF and your python code as separate files.
  - Your report should not be longer than 3 pages
  - Take into account the feedback you received on the first assignment

AFTER THIS LECTURE...

- You can explain the principles of topic modelling
- You can explain how Latent Dirichlet Allocation (LDA) works as a generative model
- You can explain the differences between extractive and abstractive summarization
- You can define the Lead-3 baseline and explain its success in benchmark data
- You can explain centrality-based summarization
- You can explain how an encoder-decoder Transformer works for abstractive summarization
- You can explain the challenges of automatic summarization
- You can define the ROUGE metric to evaluate an automatic summarizer