Natural Language Processing (NLP)

Unit 04

Pattern Matching and Topic Modeling in Python

By:

Syeda Saleha Raza

AL NAFI,
A company with a focus on education,
wellbeing and renewable energy.

اَللَّهُمَّ إِنِّي أَسْأَلُكَ عِلْمًا نَّافِعًا،

وَرِزْقًا طَيِّبًا، وَعَمَلًا مُّتَقَبَّلًا،

(O Allah, I ask You for beneficial knowledge,
goodly provision and acceptable deeds)

اے اللہ ، میں آپ سے سوال کرتی ہوں نفع بخش علم کا، طیّب رزق کا، اور اس عمل کا
جو مقبول ہو.

(Sunan Ibn Majah: 925)

Outline

- Pattern Matching using Spacy
  - Rule-based matching
  - Fuzzy matching
  - Phrase Matching
  - Code demo using spacy
- Topic Modeling
  - Latent Drichillet Allocation (LDA)
  - Non-negative Matrix Factorization
  - Code demo in using spacy and scikit-learn

Sorting Documents

Topic Modeling

- Topic modeling is a method to cluster documents and find some natural groups of items (topics).
- It provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

Latent Dirichlet Allocation (LDA)

- It is a statistical model for discovering the abstract topics.
- Each document is made up of various words, and each topic also has various words belonging to it.
- LDA is a way of automatically discovering **topics** that these sentences contain.

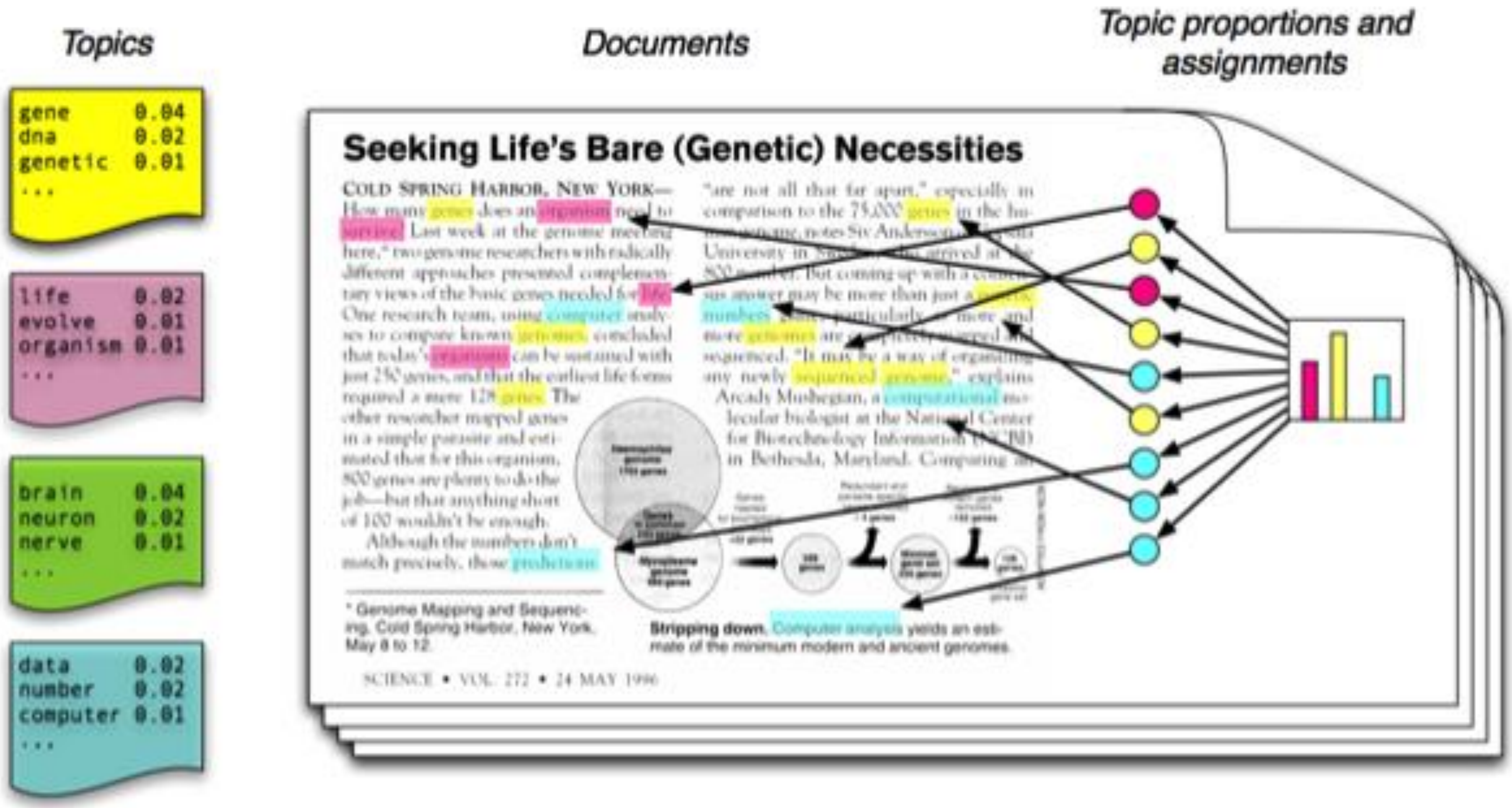


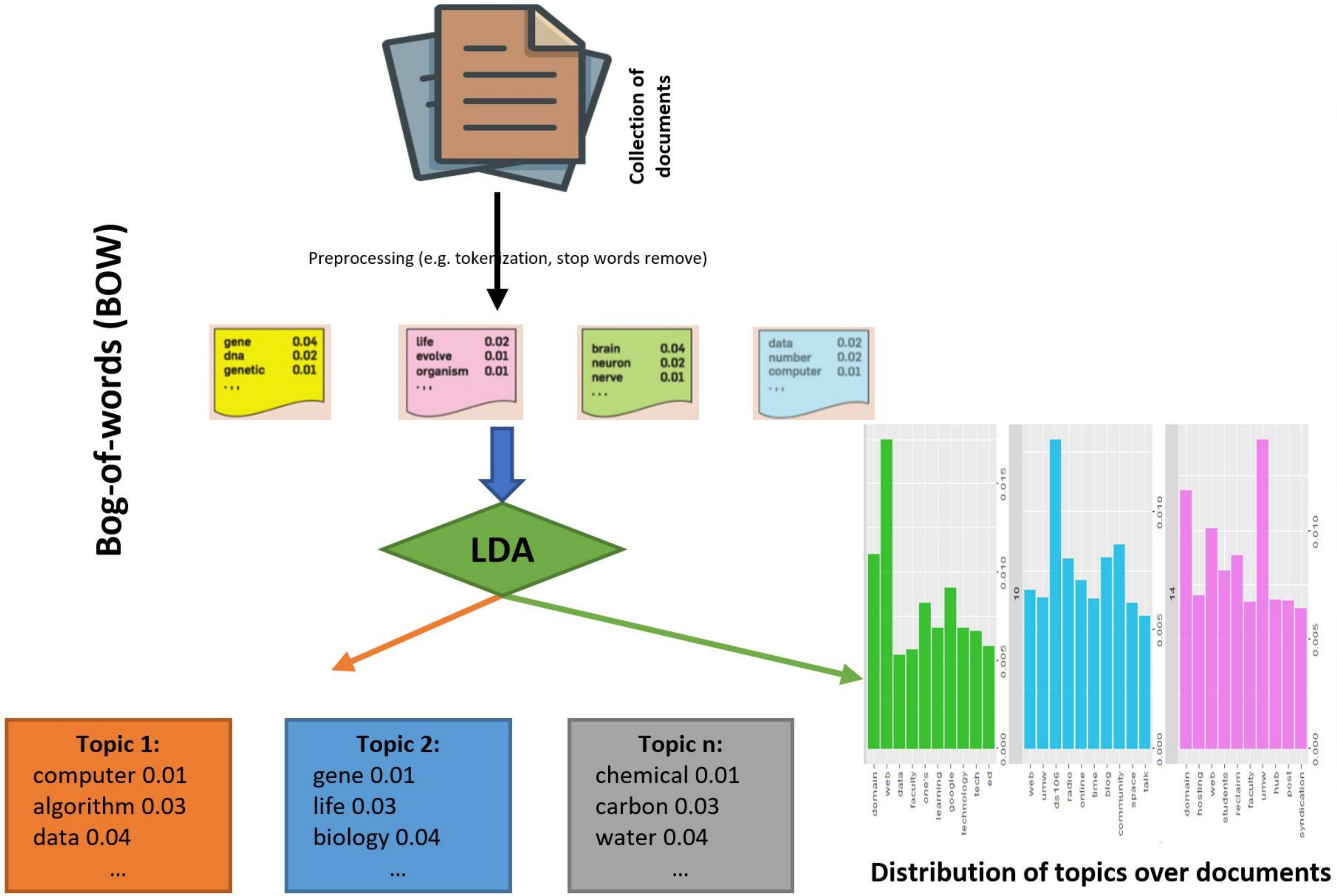
Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Topic Modelling Techniques in NLP (opengenus.org)

Example

- Suppose you have the following set of sentences:
  - I like to eat broccoli and bananas.
  - I ate a banana and spinach smoothie for breakfast.
  - Chinchillas and kittens are cute.
  - My sister adopted a kitten yesterday.
  - Look at this cute hamster munching on a piece of broccoli.
- LDA is a way of automatically discovering **topics** that these sentences contain. For example, given these sentences and asked for 2 topics,
- LDA might produce something like
  - **Sentences 1 and 2**: 100% Topic A
  - **Sentences 3 and 4**: 100% Topic B
  - **Sentence 5**: 60% Topic A, 40% Topic B
  - **Topic A**: 30% broccoli, 15% bananas, 10% breakfast, 10% munching, … (at which point, you could interpret topic A to be about food)
  - **Topic B**: 20% chinchillas, 20% kittens, 20% cute, 15% hamster, … (at which point, you could interpret topic B to be about cute animals)

How does LDA work?



Collection of documents

Preprocessing (e.g. tokenization, stop words remove)

Bog-of-words (BOW)

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

LDA

Topic 1:
computer 0.01
algorithm 0.03
data 0.04
...

Topic 2:
gene 0.01
life 0.03
biology 0.04
...

Topic n:
chemical 0.01
carbon 0.03
water 0.04
...

Distribution of topics over documents

https://www.oreilly.com/library/view/scala-machine-learning/9781788479042/b52dd4a0-3b72-43a4-b708-aeb2c2acbeb1.xhtml

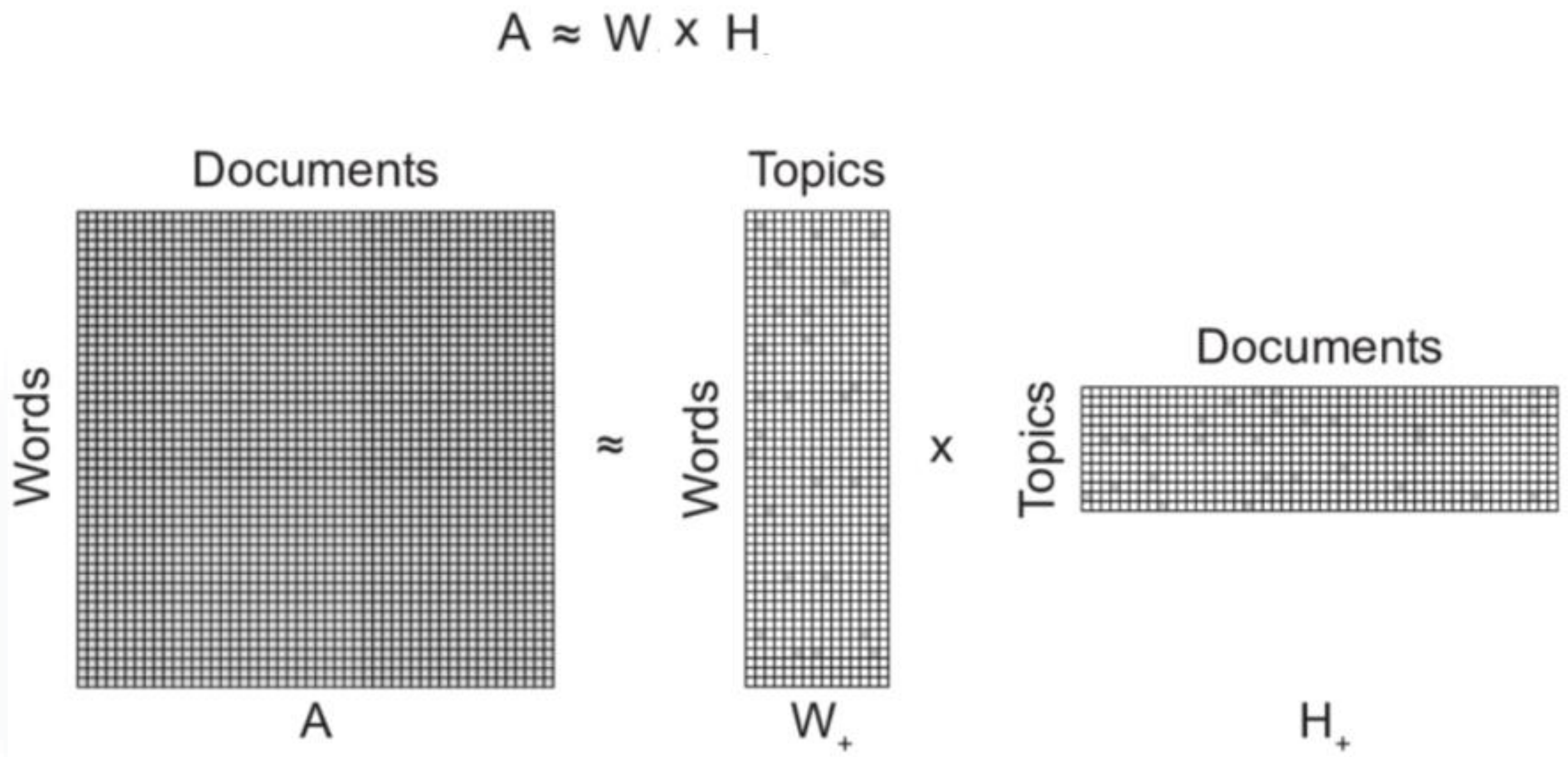
How does LDA work?

- Go through each document, and randomly assign each word in the document to one of the K topics.
- This random assignment already gives you both topic representations of all the documents and word distributions of all the topics (albeit not very good ones).
- So to improve on them, probabilities are re-estimated in an iterative process until they reach a roughly steady state where your assignments are pretty good.
- So use these assignments to estimate the topic mixtures of each document (by counting the proportion of words assigned to each topic within that document) and the words associated to each topic (by counting the proportion of words assigned to each topic overall).

Non-Negative Matrix Factorization

- Topic Modeling using Non Negative Matrix Factorization (NMF) (opengenus.org)

Matrix Factorization



Topic Modelling using NMF | Guide to Master NLP (Part 14) (analyticsvidhya.com)

Hands-on with Matrix Factorization

- Topic Modeling using Non Negative Matrix Factorization (NMF) (opengenus.org)

References

- -https://stackabuse.com/python-for-nlp-vocabulary-and-phrase-matching-with-spacy/
- https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2
- http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/
- https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
- https://www.oreilly.com/library/view/scala-machine-learning/9781788479042/b52dd4a0-3b72-43a4-b708-aeb2c2acbeb1.xhtml
- Topic Modelling using NMF | Guide to Master NLP (Part 14) (analyticsvidhya.com)

Thanks

How does LDA work?

- For each document d…
    - Go through each word w in d…
        - for each topic t, compute two things:
            - p(topic t | document d) = the proportion of words in document d that are currently assigned to topic t, and
            - p(word w | topic t) = the proportion of assignments to topic t over all documents that come from this word w.
            - Based on these probabilities, Reassign w a new topic, where topic t is chosen with probability:
                - p(topic t | document d) * p(word w | topic t)
        - In other words, in this step, we’re assuming that all topic assignments except for the current word in question are correct, and then updating the assignment of the current word using our model of how documents are generated.