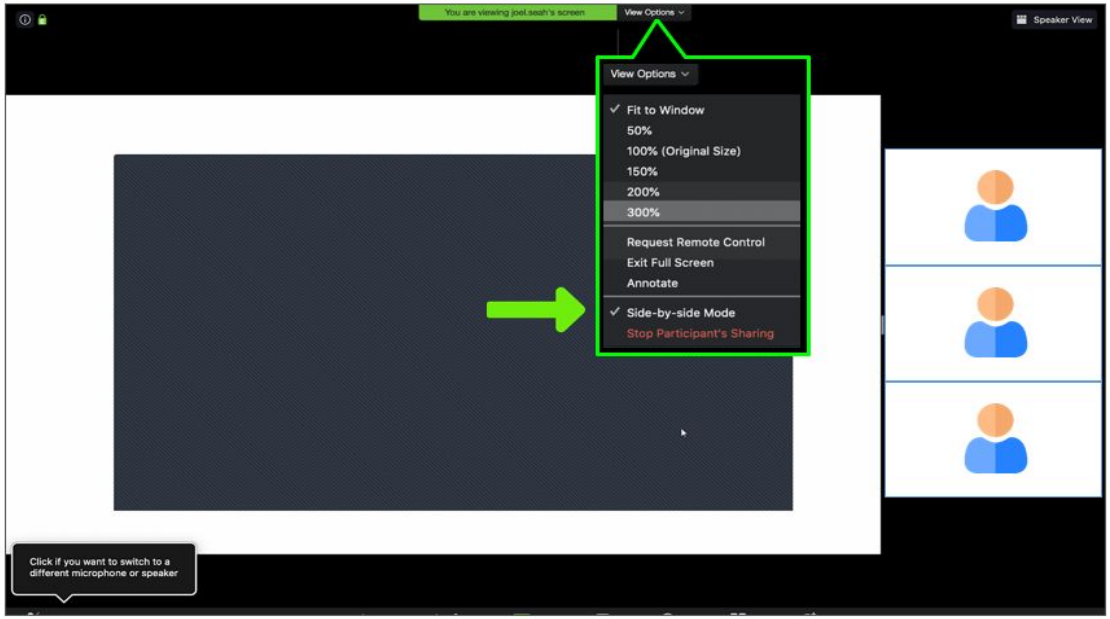
We'll Be Starting Shortly!

To help us run the workshop smoothly, kindly:

- Submit all questions using the Q&A function
- If you have an urgent request, please use the "Raise Hand" function

Using Zoom: Viewing Mode



Side-By-Side Mode

- When sharing screen (slide share)
- With small thumbnails of people on the sidebar

STEPS:

1. View Options
2. Side-By-Side Mode

Topic Modeling and Sentiment Analysis

Albert ‘Bash’ Yumol

Eskwelabs x Shopee Code League

Bash Yumol

Introduction

Who am I?

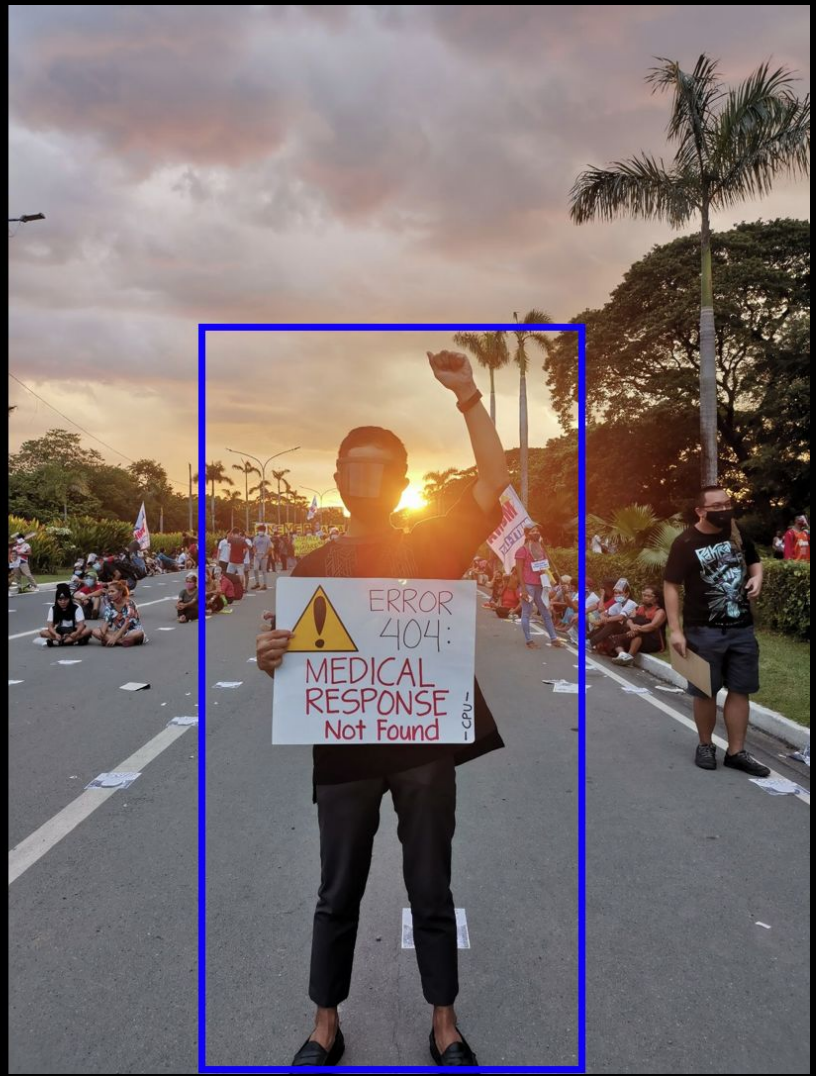
Name: Albert 'Bash' Yumol

Lives: Manila Philippines

Interests: Physics, AI, Big Data,
Cryptography, IoT, Activism

Occupation: Data Scientist and AI Consultant, EdTech

Connect: https://www.linkedin.com/in/albertyumol/
https://github.com/albertyumol
https://albertyumol.github.io/



Mission

Eskwelabs is an online data upskilling school based in the Philippines driving social mobility in the future of work.

We build data skills for workers and teams through **mentor-led project-based upskilling**.

For Individuals

90% job placement within 90 days
50% increase in income

Source: Job outcome survey reported by students from Cohort I-III, Eskwelabs Data Science program.

For Companies

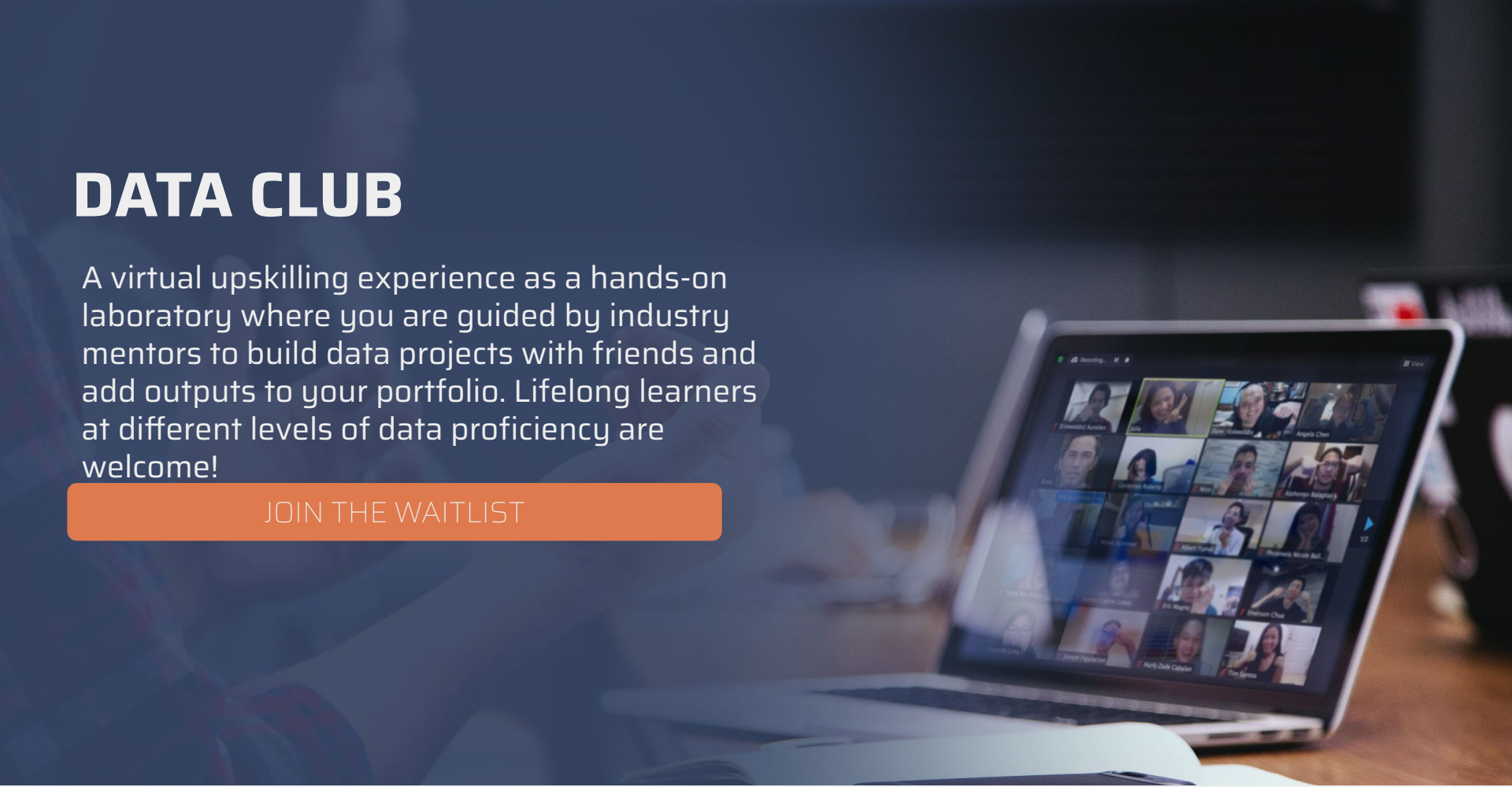
Build or buy talent
Pivot to mid to high-value work

TaskUs · UnionBank · pwc · MSCI · Uber · InterContinental Hotels & Resorts · Shopee · kumu · The World Bank · AboitizPower

DATA CLUB

A virtual upskilling experience as a hands-on laboratory where you are guided by industry mentors to build data projects with friends and add outputs to your portfolio. Lifelong learners at different levels of data proficiency are welcome!

JOIN THE WAITLIST



SPRINT TOPICS

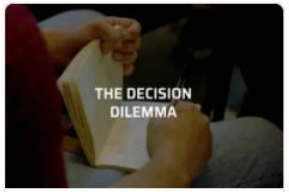
www.eskwelabs.com/data-club

Interactive Data Visualization with PowerBI

Learn how to turn visualization into insights with one of the most powerful tools for data analysis - PowerBI - while building a beautiful, and interactive dashboard to track the latest pandemic developments.

Beginner · No Code · PowerBI

READ MORE →



What-If Analysis and Optimization with Solver in Excel

Make better everyday and business decisions using Excel Solver that optimize allocation of resources.

Beginner · No Code

READ MORE →

Create Exploratory Data Analysis in Python

Tell a story through data on how far the world has progressed on the UN's Global Goals, a universal call to action to end poverty, protect the planet, and ensure that all people enjoy peace and prosperity by 2030.

Data for Good · Beginner · Python

READ MORE →

Save Time by Automating Work in Excel

Ever wonder how some people manage to get their work done faster? Their secret is working smart by using Excel VBA to automate repeatable tasks.

Beginner · No Code

READ MORE →

Introduction to Object Recognition

Learn the foundations of computer vision and implement your own object detection algorithm and identify an object of your choice.

Intermediate · Python

READ MORE →

The Rise of BTS

Create a bar chart race using Python to visualize how music artist popularity changed over time.

Beginner · Python · Data Viz

READ MORE →

Design Data Strategy for a Fast Food Restaurant

Help craft the data strategy for your favourite fast food chain by understanding how data can serve business goals.

Beginner · No Code

READ MORE →

Data meets Don Draper - Customer Segmentation Analysis

The digital economy means customers are online. Help a creative ad agency target the right audiences with digital marketing.

Beginner · No Code · SQL

READ MORE →

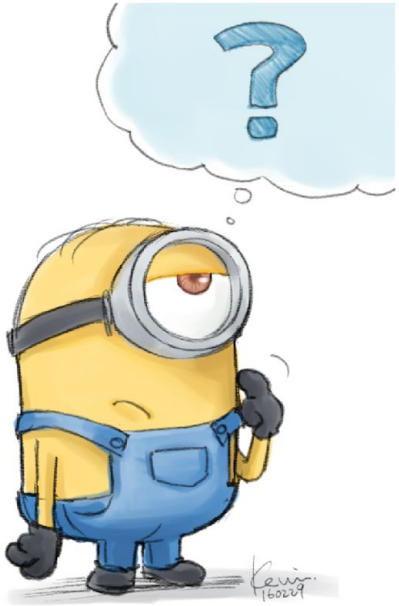
RECAP

Natural Language Processing

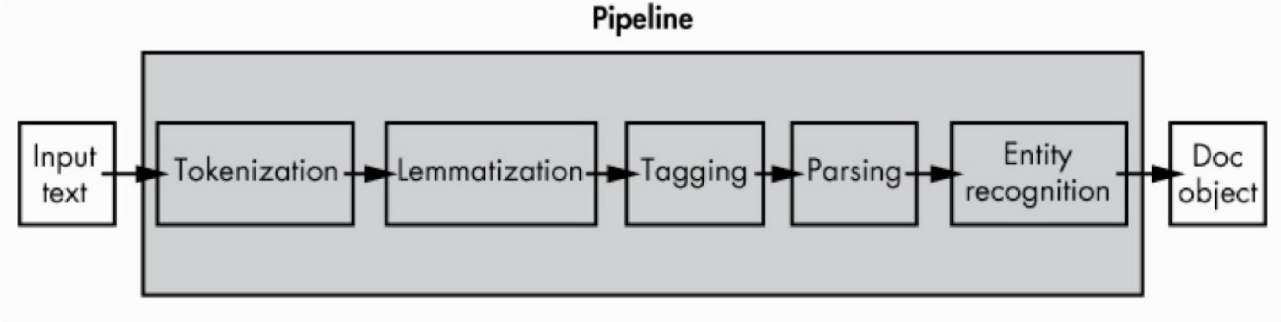
a branch of artificial intelligence that helps machine understand and respond to human language.

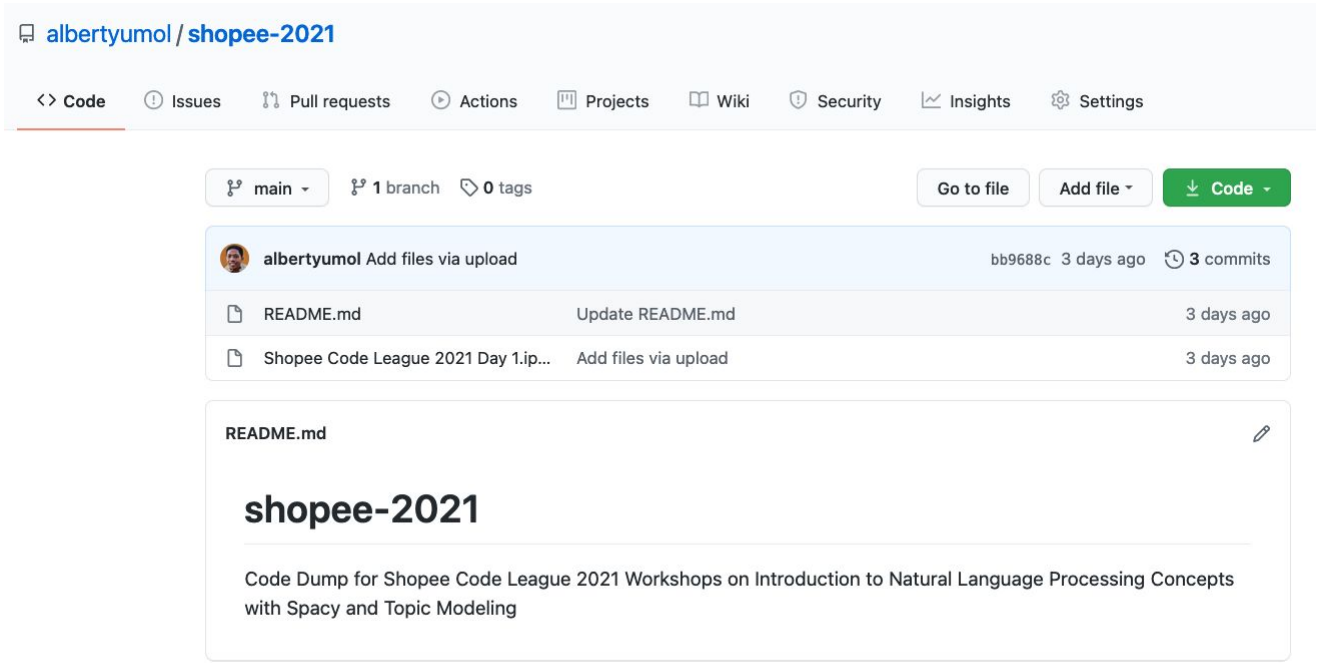
Why NLP is Hard?

1. Ambiguity
2. Scale
3. Sparsity
4. Variation
5. Expressivity
6. Unmodeled Variables
7. Unknown representations



Basic NLP Operations with SpaCy




albertyumol / shopee-2021
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
main
1 branch
0 tags
Go to file
Add file
Code
albertyumol Add files via upload
bb9688c 3 days ago
3 commits
README.md
Update README.md
3 days ago
Shopee Code League 2021 Day 1.ip...
Add files via upload
3 days ago
README.md
shopee-2021
Code Dump for Shopee Code League 2021 Workshops on Introduction to Natural Language Processing Concepts with Spacy and Topic Modeling

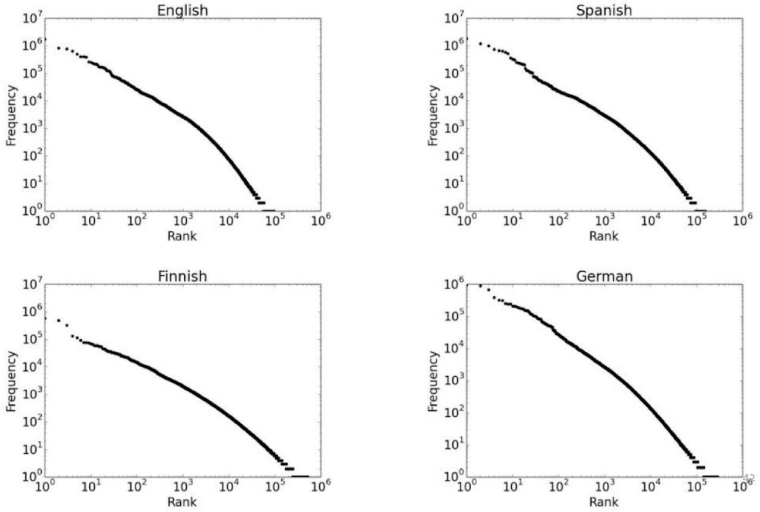
Start

Objectives

TF-IDF

Sparsity

- Regardless of how large our corpus is, there will be a lot of infrequent words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen

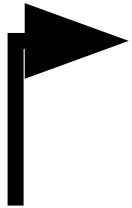


TF-IDF

weighted **frequency** of the word in each document

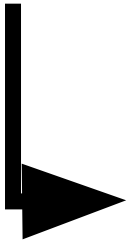
weight of **rare words** across all documents

TF-IDF score



total number of documents

number of occurrences of *i* in *j*



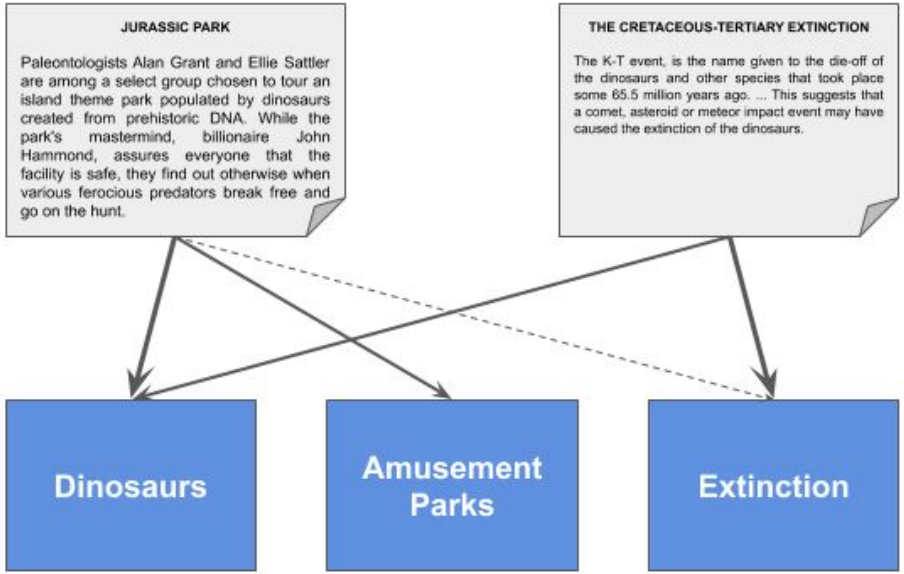
number of documents containing *i*

Code Time

Sentiment Analysis with VADER

Code Time

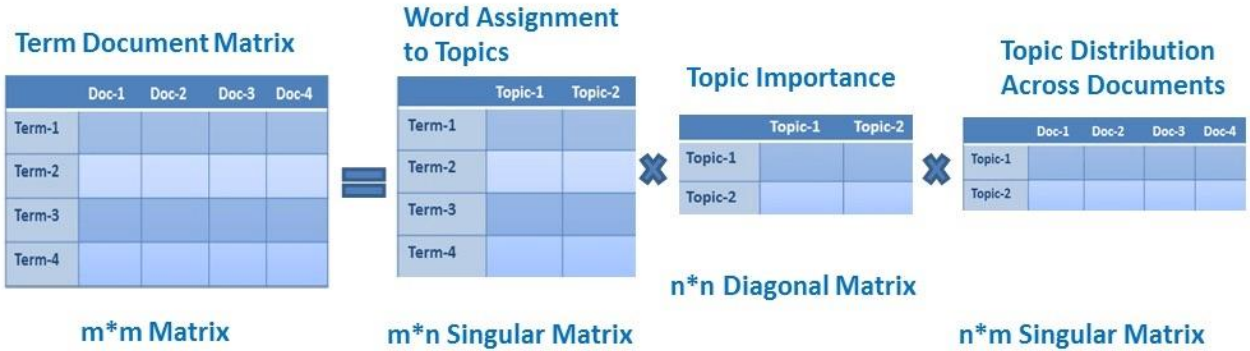
Topic Modeling



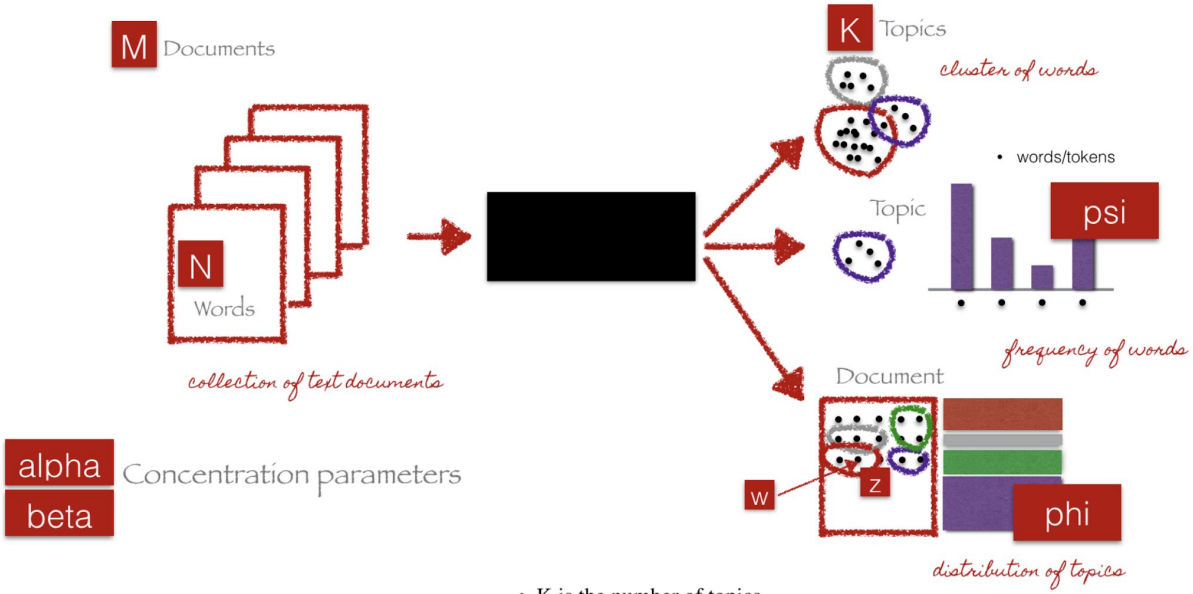
Latent Semantic Indexing (LSI)

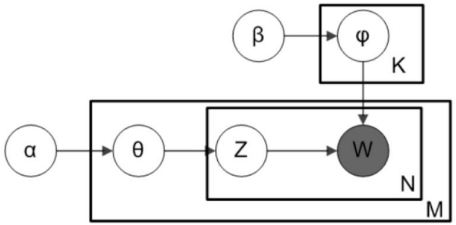
$$
\begin{array}{ccccc}
 & & X & & \\
 & & (\mathbf{d}_j) & & \\
 & & \downarrow & & \\
(\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,j} & \cdots & x_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,j} & \cdots & x_{m,n} \end{bmatrix} & = & (\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} \cdots \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix}
\end{array}
\begin{array}{c}
U \\
\\
\begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_l \end{bmatrix}
\end{array}
\begin{array}{c}
V^T \\
(\mathbf{d}_j) \\
\downarrow \\
\begin{bmatrix} | & & | \\ \mathbf{v}_1 & & \mathbf{v}_l \\ | & & | \end{bmatrix}
\end{array}
$$

Latent Semantic Indexing (LSI)



Latent Dirichlet Allocation (LDA)

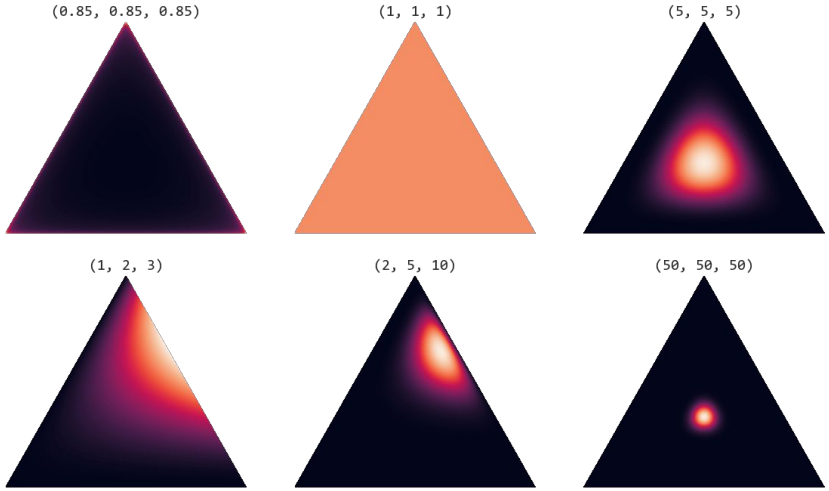




- K is the number of topics
- N is the number of words in the document
- M is the number of documents to analyse
- α is the Dirichlet-prior concentration parameter of the per-document topic distribution
- β is the same parameter of the per-topic word distribution
- ϕ(k) is the word distribution for topic k
- θ(i) is the topic distribution for document i
- z(i,j) is the topic assignment for w(i,j)
- w(i,j) is the j-th word in the i-th document
- ϕ and θ are Dirichlet distributions, z and w are multinomials.

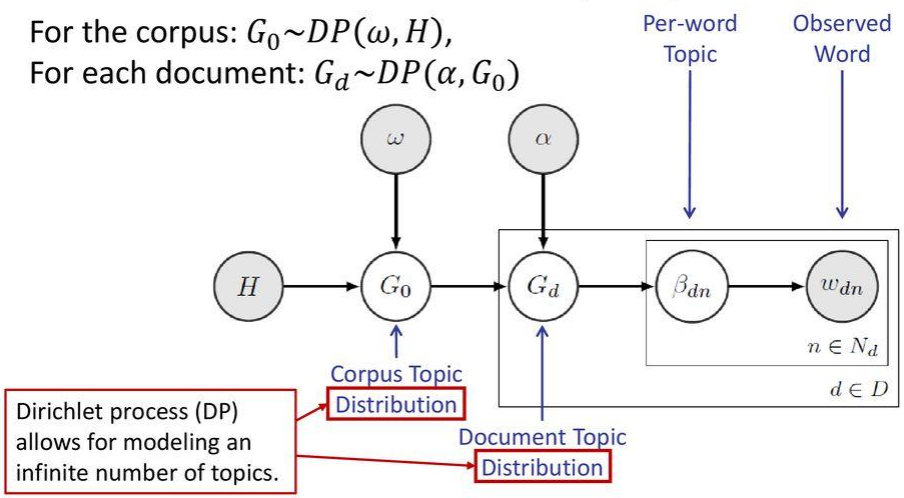
Dirichlet Distribution

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{\mathrm{Beta}(\boldsymbol{\alpha})} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}, \text{ where } \mathrm{Beta}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \text{ and } \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k).$$



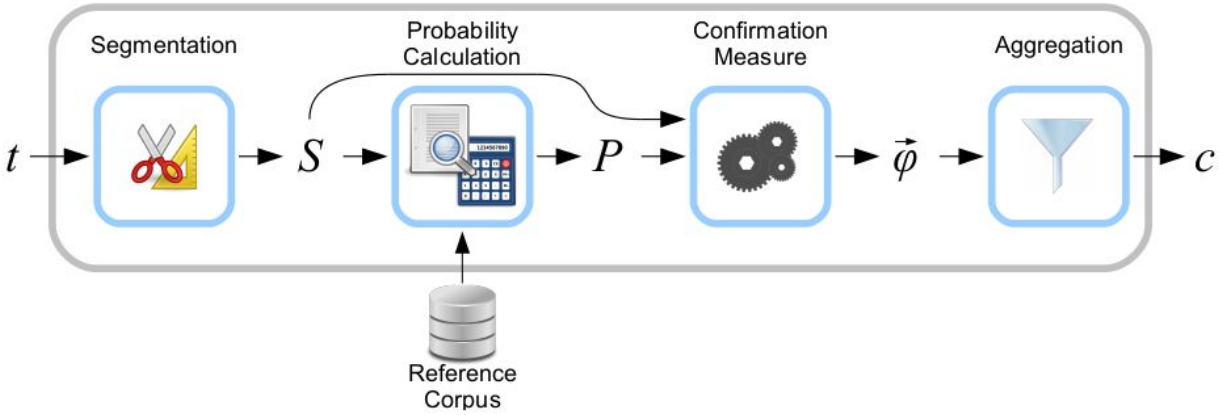
Hierarchical Dirichlet Process (HDP)

For the corpus: $G_0 \sim DP(\omega, H)$,
For each document: $G_d \sim DP(\alpha, G_0)$



H: base topic distribution (e.g., Dirichlet distribution Dir($\boldsymbol{\alpha}$));
ω: corpus topic concentration parameter; α: document topic concentration parameter

Topic Coherence



Code Time

Code Time

References

https://web.stanford.edu/~jurafsky/slp3/

https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf

Natural Language Processing with Python and Spacy: Yuli Vasiliev

Resources

https://course.spacy.io/en/

https://www.nltk.org/book/

https://datasets.quantumstat.com/

https://notebooks.quantumstat.com/

https://radimrehurek.com/gensim/auto_examples/

THANK YOU!

ESKWELABS

Your Feedback Matters!

bit.ly/3hmJ3Nr