PROJECT

LITERARY ANALYSIS WITH NLP

TOPIC MODELING

DATE

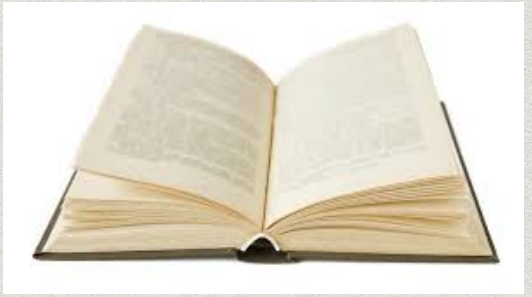
PYTEXAS

CLIENT

RACHEL BRYNSVOLD

Roadmap

- Intro to NLP, LDA
- Vector Representations of Books



```
[(41, 1),
 (189, 3),
 (441, 2),
 (1368, 1), …]
```

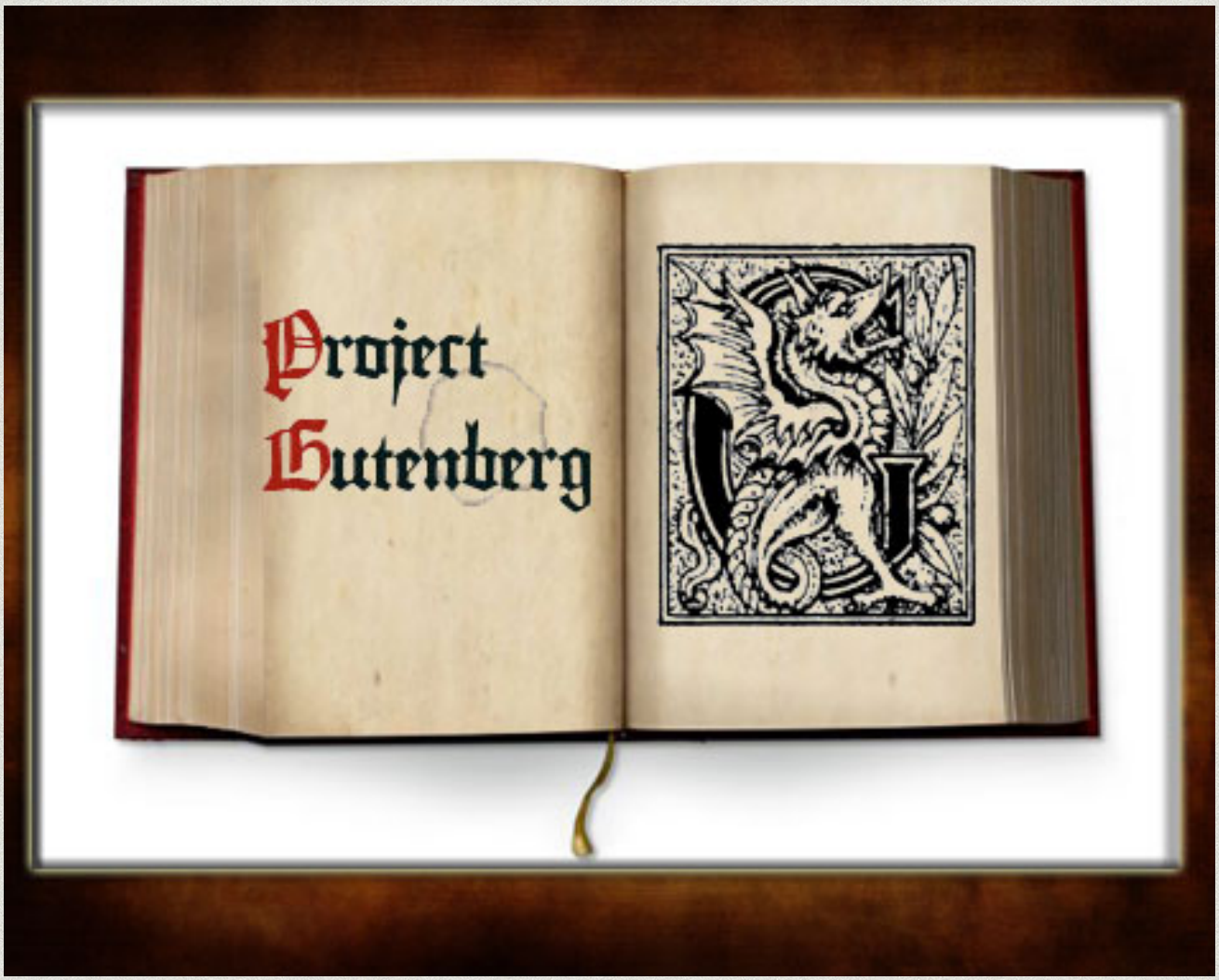
- Demo!

NLP Primer

- **NLP**: Natural Language Processing
- **Token**: An individual word (string format)
- **Document**: A single text
- **Corpus**: Collection (‘body’) of documents being analyzed
- **Stop Words**: Common words that don’t contribute to document meaning
- **Bag of Words**: Simplifying representation that omits grammar and word order and tracks word occurrence

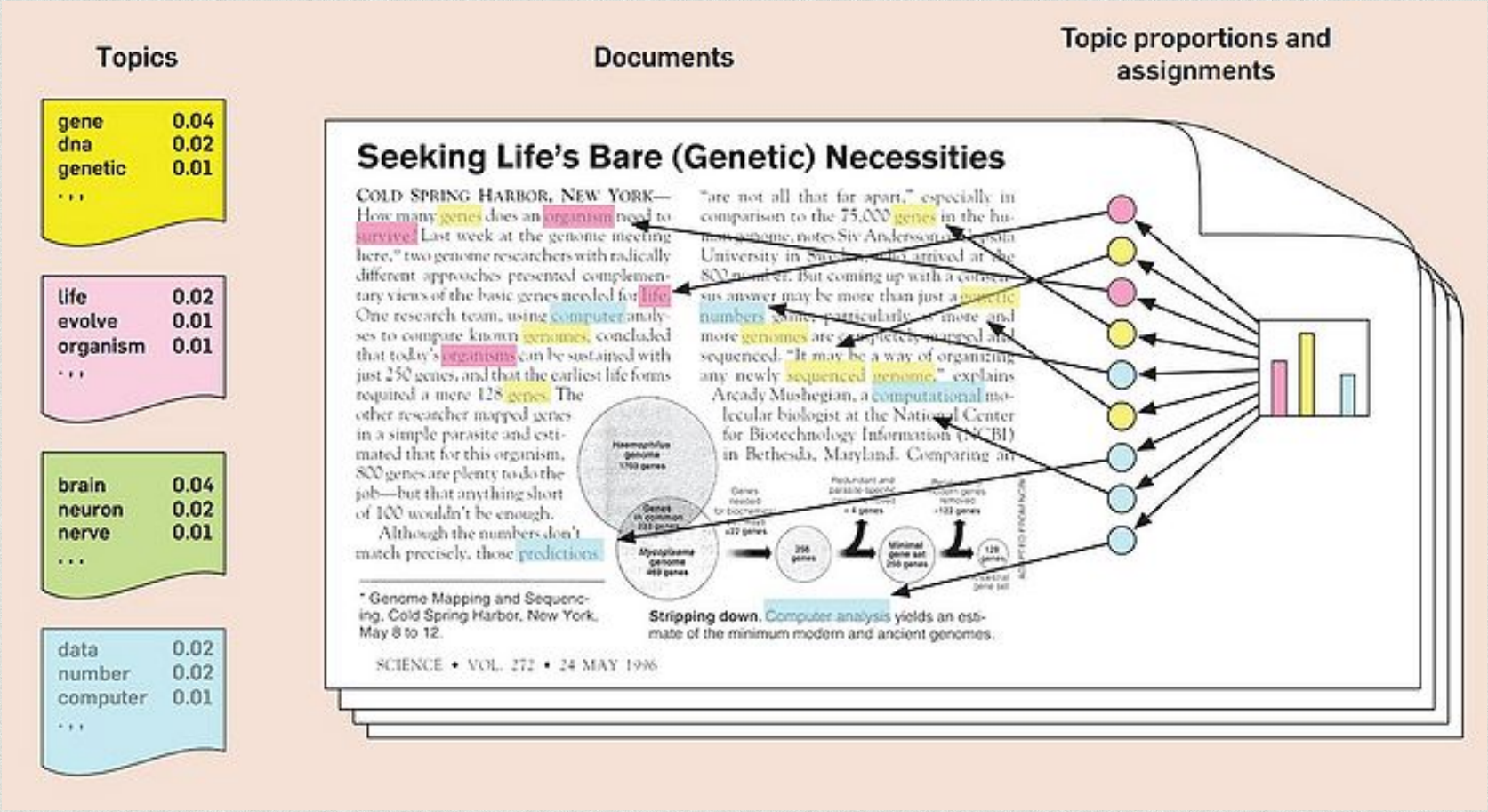
Data: Project Gutenberg

- Repository of ebooks
- 28,000-text *corpus* modeled
- ~10GB

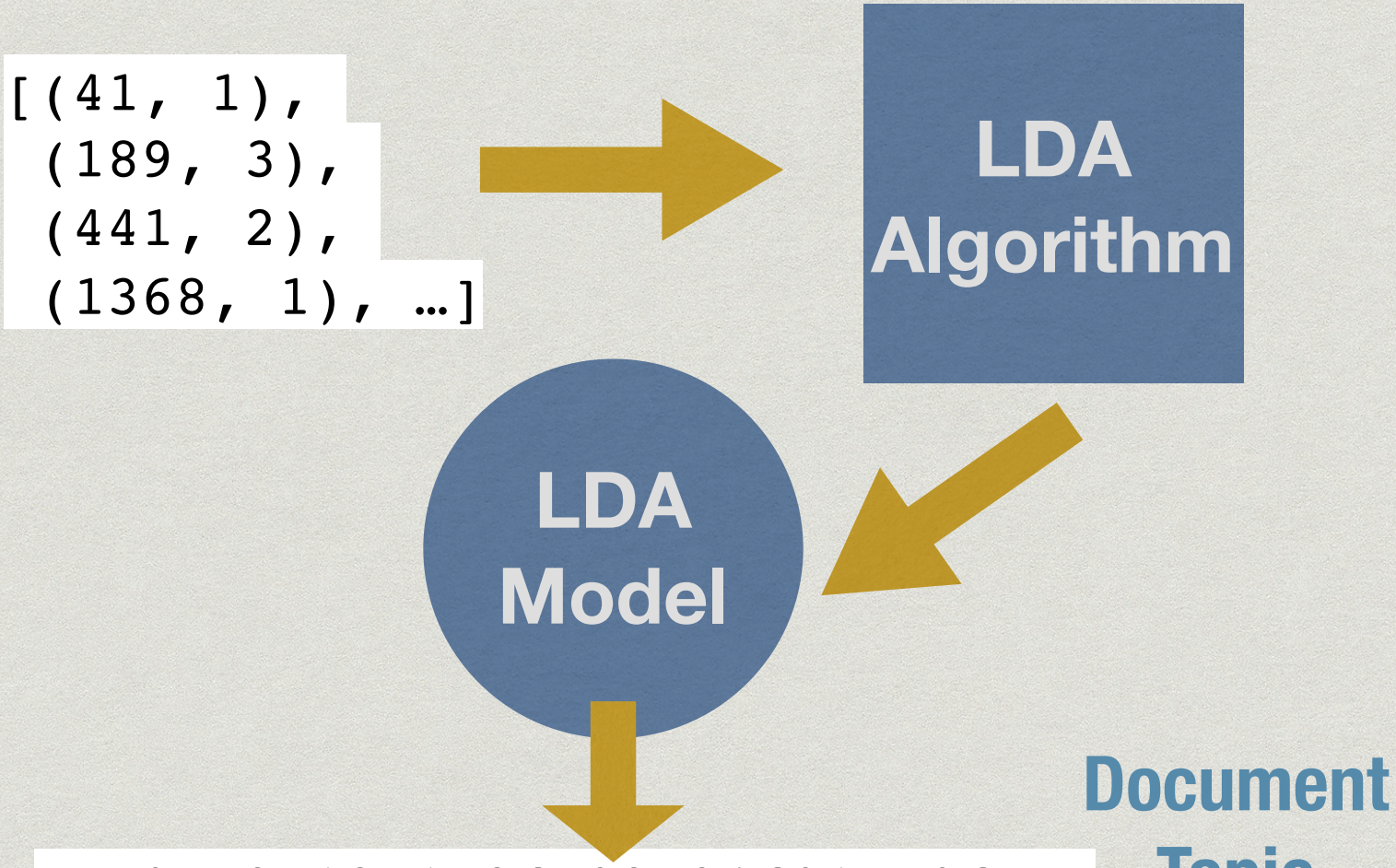


LDA - *Latent Dirichlet Allocation*

- Unsupervised learning method
- Intermediate layer between words and documents



LDA Take-Aways



Enabling Technologies

Enabling Tech: Rambo

- Automates configuration, provisioning of vm's
- Multiple platforms supported

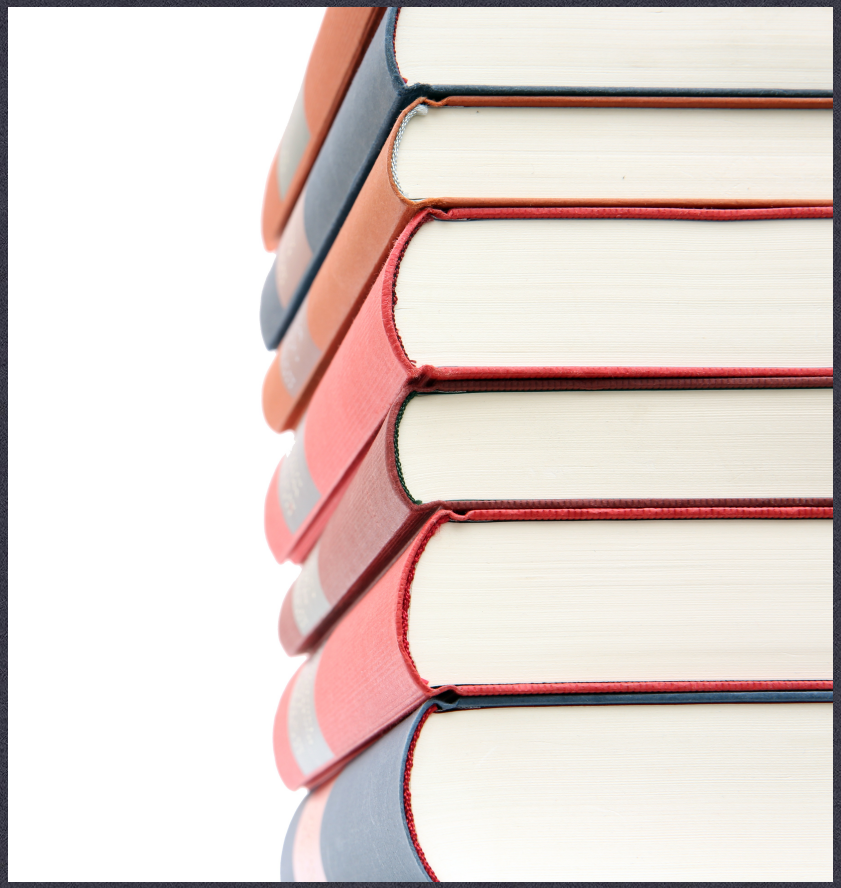
- **Makes data science work reproducible!**

DEMO TIME

DON'T TAKE MY WORD FOR IT

TRY IT FOR YOURSELF!

GITHUB.COM/
RBRYNSVOLD/CAPSTONE



THANK YOU

RBRYNSVOLD

RACHELADELEB

RACHELBRYNSVOLD

BACKUP

Next Steps

- Add and tune text pre-processing
- Model tuning/optimization
- Reproducibility (ongoing improvements)

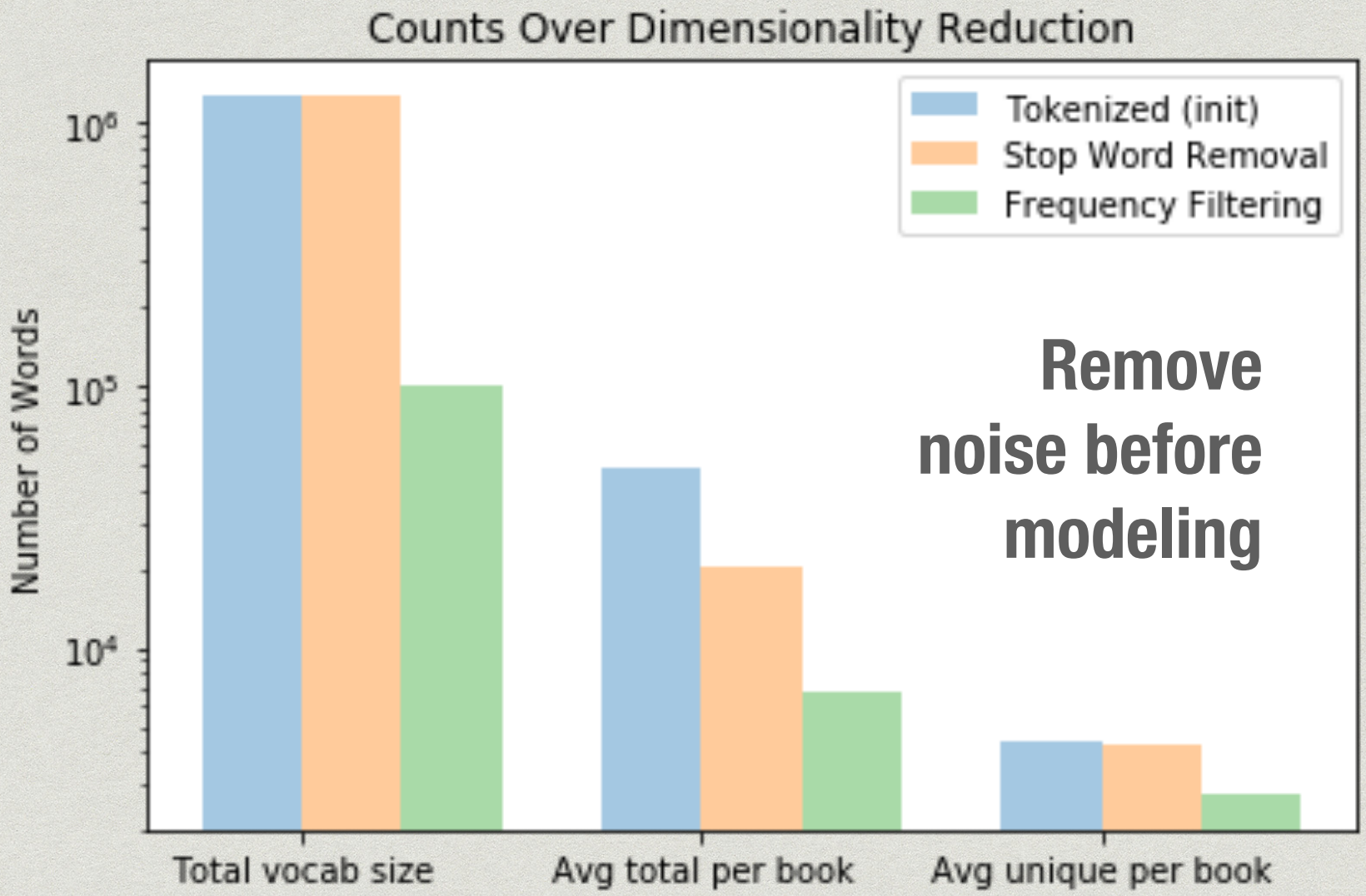
“When we try to pick out anything by itself, we find it hitched to everything else in the Universe.”

–John Muir

Tools

ANACONDA®

Dimensionality Reduction of Corpus



Results: Topic Terms

Topic 18: "Chemistry"

```
"acid" + "solution" +
"oil" + "temperature"
+ "gas" + "chemical"
+"liquid" + "mixture"
+ "alcohol" +
"sulphuric"
```

Topic 8: "Rome"

```
"rome" + "greek"
+"italian" +"di" +
"pope" + "italy" +
"romans" + "temple" +
"caesar" + ”emperor"
```

Topic 42: "Classical Music"

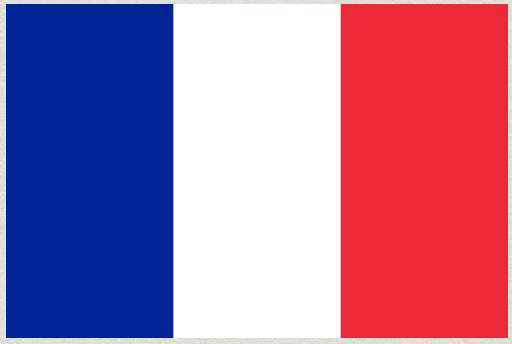
```
"musical" + "paul" + "von" + "opera" + "der" + "und"
+  "composer" + "piano" + "violin" + "italian"
```

Results: Term Topics

“France”

"et" + "le" + "des" + "les" + "du" + "que" + "en" + "qui" + "il" + ”un”

"madame" + "duke" + "louis" + "paris" + "majesty" + "monsieur" + "france" + "princess" + "honour" + ”le”



"km" + "billion" + "na" + "million" + "population" + "islands" + "december" + "comparison" + “president" + "economic"

Challenges → Solutions

Memory

Streaming Data

Computational Intensity

Distributed Computing

ETL Steps

-Download data from gutenberg (wget)
-Consolidate file structure
-Set up mirror, mirroring script
-Unzip, Clean (remove duplicates, remove
-Gutenberg headers)

ETL = Extract/Transform/Load