Comparative Vector Space Modeling of Tolkien's Works

Jeffrey R. Tharsen

University of Chicago

MiddleMoot 2021
October 9, 2021

Mae g'ovannen!

Part 1 :

Digital Philology: Sources and Methods

Part 2 :

Word Frequencies in Tolkien's Primary Works

Part 3:

Vector Space Models of Tolkien's Primary Works

Github link:

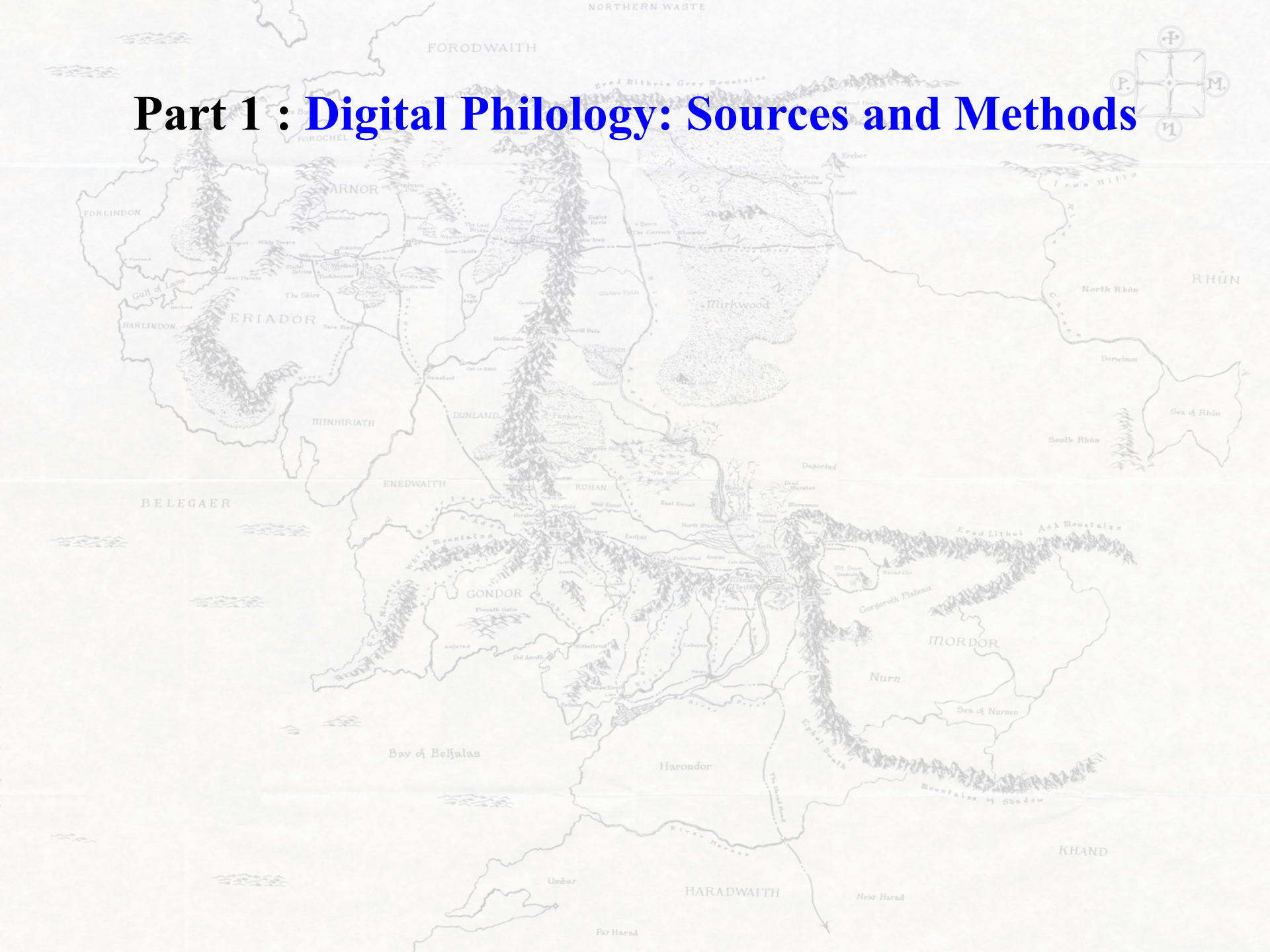
github.com/thars3n/Tolkien_Vectors

Today's Big Questions :

When we use machines (computers) to model languages and literature, what can they tell us?

Is there anything there that we didn't already know?

Part 1 : Digital Philology: Sources and Methods



Part 1 : Digital Philology: Sources and Methods

- We always start with the books (sources)
  - *The Hobbit*
  - *The Lord of the Rings* (3 volumes)
  - *The Silmarillion*
  - *The Lost Tales* (2 volumes)

Part 1 : Digital Philology: Sources and Methods

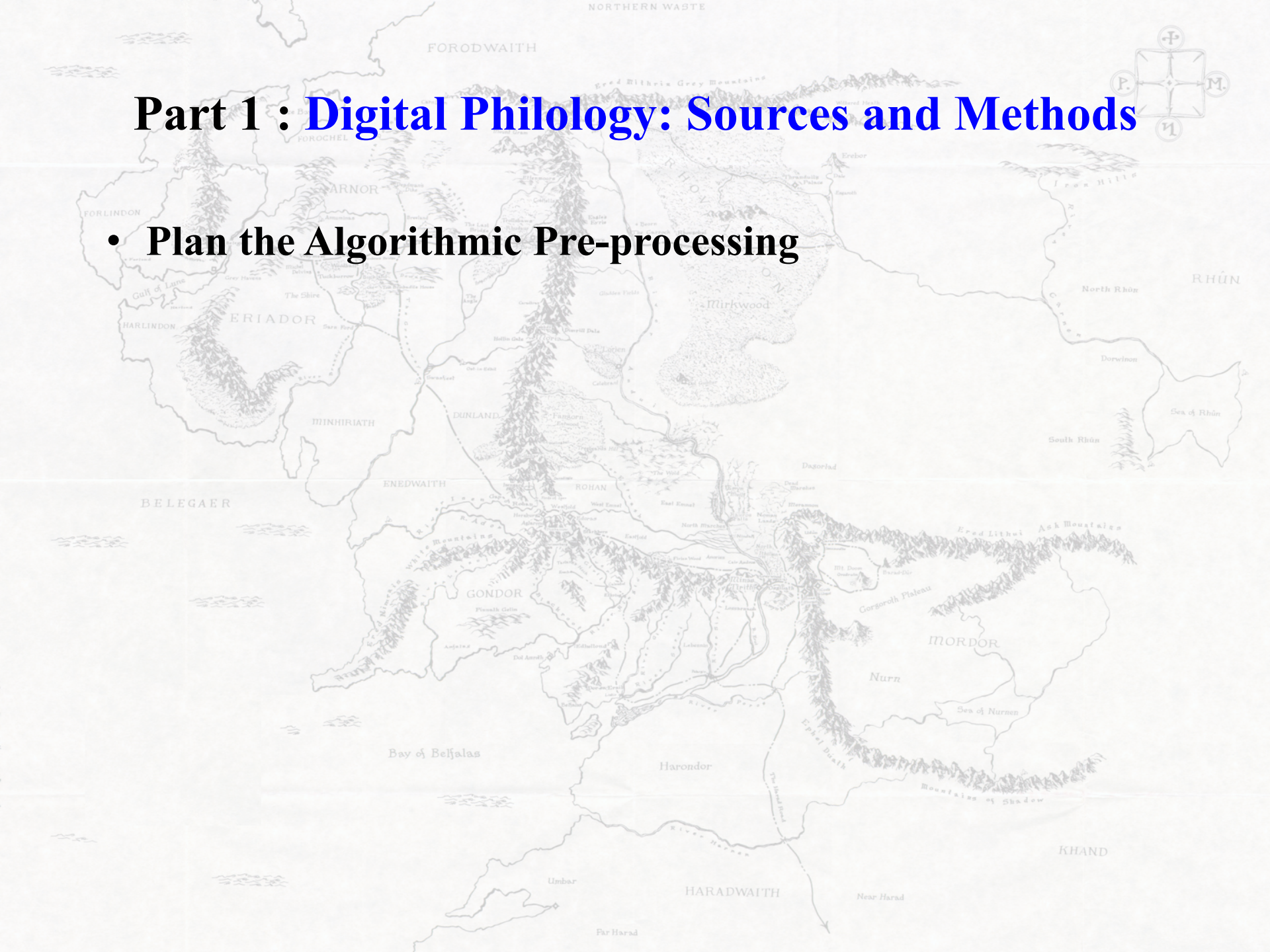
- We start with the books (sources)
  - *The Hobbit*
  - *The Lord of the Rings* (3 volumes)
  - *The Silmarillion* (ed. Christopher Tolkien)
  - *The Lost Tales* (2 volumes) (ed. Christopher Tolkien)

Part 1 : Digital Philology: Sources and Methods

- We start with the books (sources)
  - *The Hobbit*
  - *The Lord of the Rings* (3 volumes)
  - *The Silmarillion* (ed. Christopher Tolkien)
  - *The Lost Tales* (2 volumes) (ed. Christopher Tolkien)
- The sources will almost always need pre-processing
  - Check for errors in the text (there were some: é = j etc)
  - Remove "non-head-text"/"extraneous" sections (*LT* Notes & Commentary, Front & Back Matter)
  - Decide what to keep and what to take out

Part 1 : Digital Philology: Sources and Methods

- **Plan the Algorithmic Pre-processing**



Part 1 : Digital Philology: Sources and Methods

- Plan the Algorithmic Pre-processing
  - For the Frequencies:
    - Remove punctuation and white space
    - Remove “stop words” (common words)

Part 1 : Digital Philology: Sources and Methods

- Plan the Algorithmic Pre-processing
  - For the Frequencies:
    - Remove punctuation and white space
    - Remove “stop words” (common words)
  - For the Vector Space (Word2vec) Models:
    - Cut into sentences
    - Remove punctuation and white space

Part 2 : Word Frequencies in Tolkien's Primary Works

- **The script (code) is in the "code" folder**
  - **Python .ipynb file :**
    "SpaCy Most Frequent Words and Lemmas.ipynb"

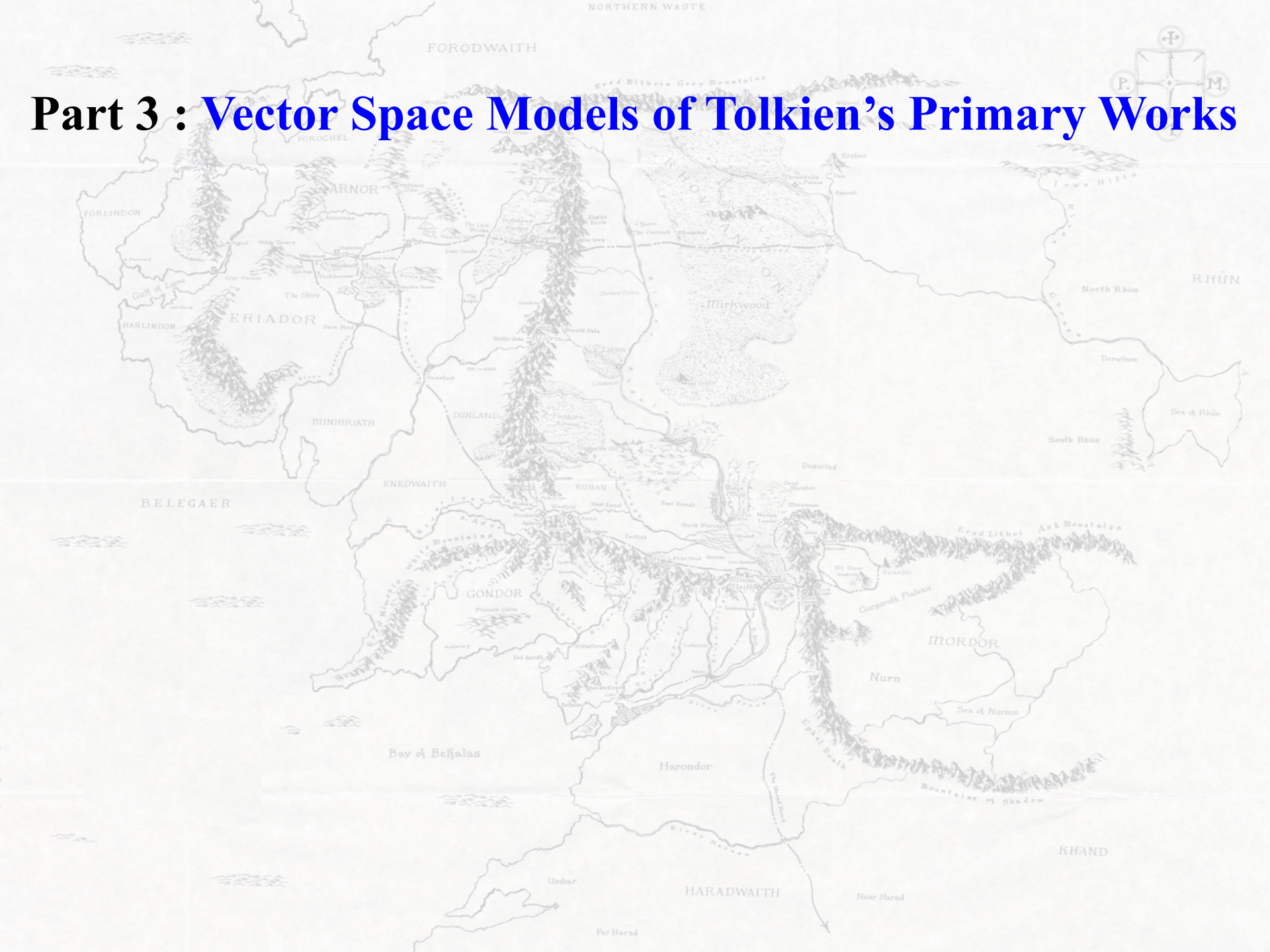
Part 2 : Word Frequencies in Tolkien's Primary Works

- The script (code) is in the "code" folder
  - Python .ipynb file :
    "SpaCy Most Frequent Words and Lemmas.ipynb"
- The output (100 most frequent words and lemmas) for each *oeuvre* (volumes aggregated) is in the "most_common_words_and_lemmas" folder

Part 2 : Word Frequencies in Tolkien's Primary Works

- **The script (code) is in the "code" folder**
  - **Python .ipynb file :**
    "SpaCy Most Frequent Words and Lemmas.ipynb"
- **The output (100 most frequent words and lemmas) for each *oeuvre* (volumes aggregated) is in the "most_common_words_and_lemmas" folder**
- ***What are lemmas? Why do we care?***

Part 3 : Vector Space Models of Tolkien's Primary Works



Part 3 : Vector Space Models of Tolkien's Primary Works

- ***What is a "vector space model"? Why do we care?***
- ***What can it tell us about the words/phrases Tolkien used?***
- **Word2vec : "proximate semantics"**
  = semantic similarity
  = <u>word associations</u>

Part 3 : Vector Space Models of Tolkien's Primary Works

- ***What is a "vector space model"? Why do we care?***
- ***What can it tell us about the words/phrases Tolkien used?***
- **Word2vec : "proximate semantics"**
  = semantic similarity
  = word associations
- ***How many models should we build?***
  (1 each, and then "one model to rule them all")
- ***Will they all work equally well?***
  (What should our Word2vec parameters be?)

Part 3 : Vector Space Models of Tolkien's Primary Works

- See the Results – in the code (1 word/phrase at a time)
  - Python .ipynb file for building the models:
    "Word2Vec Models for Tolkien.ipynb"

Part 3 : Vector Space Models of Tolkien's Primary Works

- **See the Results – in the code (1 word/phrase at a time)**
  - **Python .ipynb file for building the models: "Word2Vec Models for Tolkien.ipynb"**

But far more fun is to use the

- **Interactive Visualization of Vector Space models:**

https://projector.tensorflow.org/

- **All model files (vectors and metadata/words) are in the "projector_files" directory**

Thank you!

Hantanyel órenyallo!

Github link:

github.com/thars3n/Tolkien_Vectors