Parsing “War and Peace”

David Chen

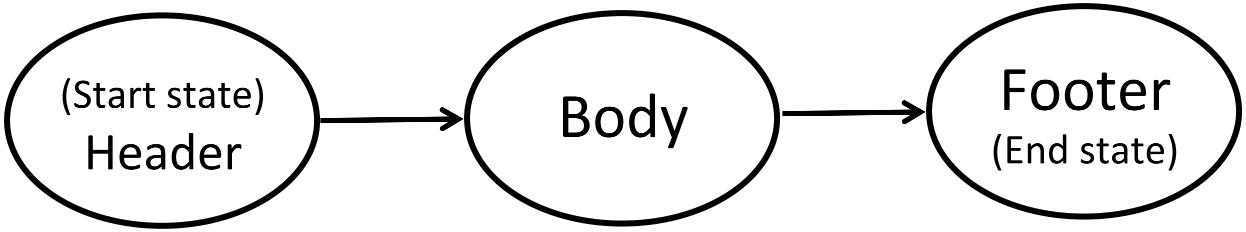
12/22/2018

Assignment: Parse Tolstoy’s War and Peace into nested format including book|chapter|paragraph|sentence|words

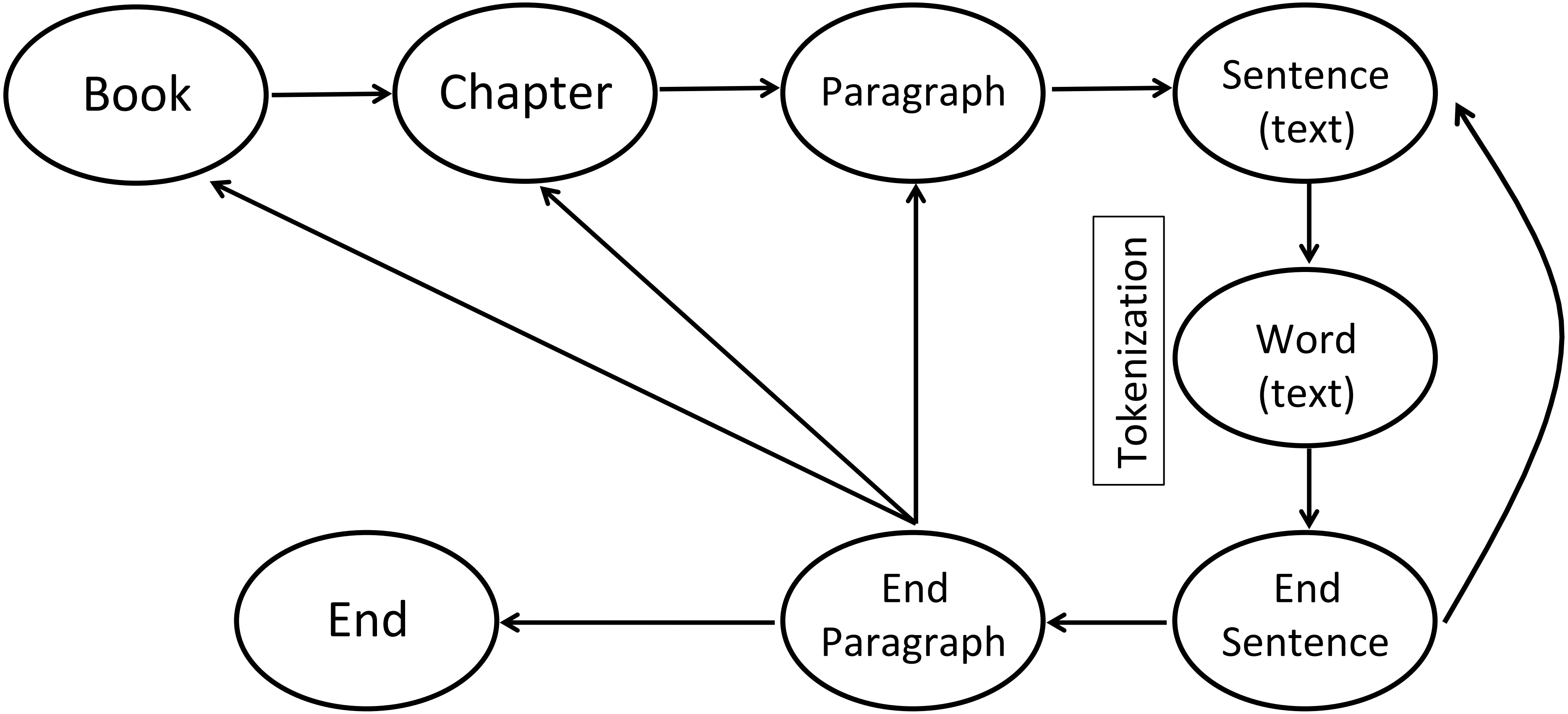
Approach: Used a state machine framework for organizing the problem

Container: nested dictionary
Seralization: JSON

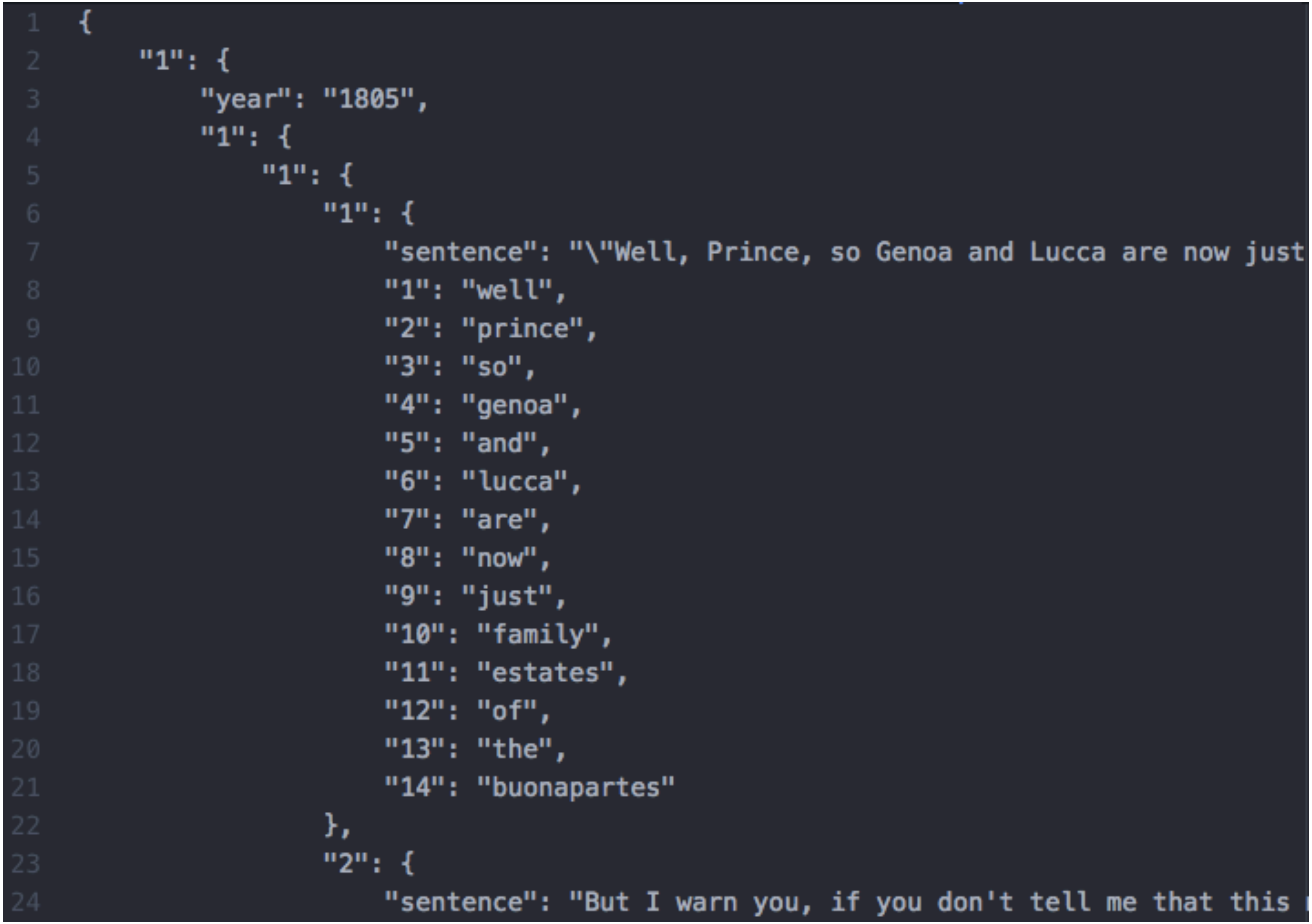
Header|body parser State Diagram:



book|chapter|paragraph|sentence|word parser
State Diagram:



Nested dictionary structure (resulting JSON file)



```
{
    "1": {
        "year": "1805",
        "1": {
            "1": {
                "1": {
                    "sentence": "\"Well, Prince, so Genoa and Lucca are now just
                    "1": "well",
                    "2": "prince",
                    "3": "so",
                    "4": "genoa",
                    "5": "and",
                    "6": "lucca",
                    "7": "are",
                    "8": "now",
                    "9": "just",
                    "10": "family",
                    "11": "estates",
                    "12": "of",
                    "13": "the",
                    "14": "buonapartes"
                },
                "2": {
                    "sentence": "But I warn you, if you don't tell me that this
```

Sentence and word tokenization

Need to include edge cases:

- **Sentences** include honorifics/titles that look like periods: Mr., Mrs., Dr.
- **Words** have hyphens: “But before Pierre—who at that moment imagined himself to be Napoleon… and captured London—could pronounce…” vs. “he saw a well-built and handsome young officer”
  - Solution: separate words in both cases. ‘Pierre’, ‘who’, and ‘well’, ‘built’ by replacing ‘-’ with ‘, ’

My approach: Use NLTK library with sent_tokenize and word_tokenize. Create separate filters for punctuation and hyphens.

Result (by character count):

Ratio of direct-to-indirect speech = 0.0726

Total % of direct speech = 6.76 %

Total % of indirect speech = 93.24 %

All text analysis located in ‘Text_analysis.ipynb’

Book & chapter length visualization:

Books 10 and 11 are outliers in number of chapters and length

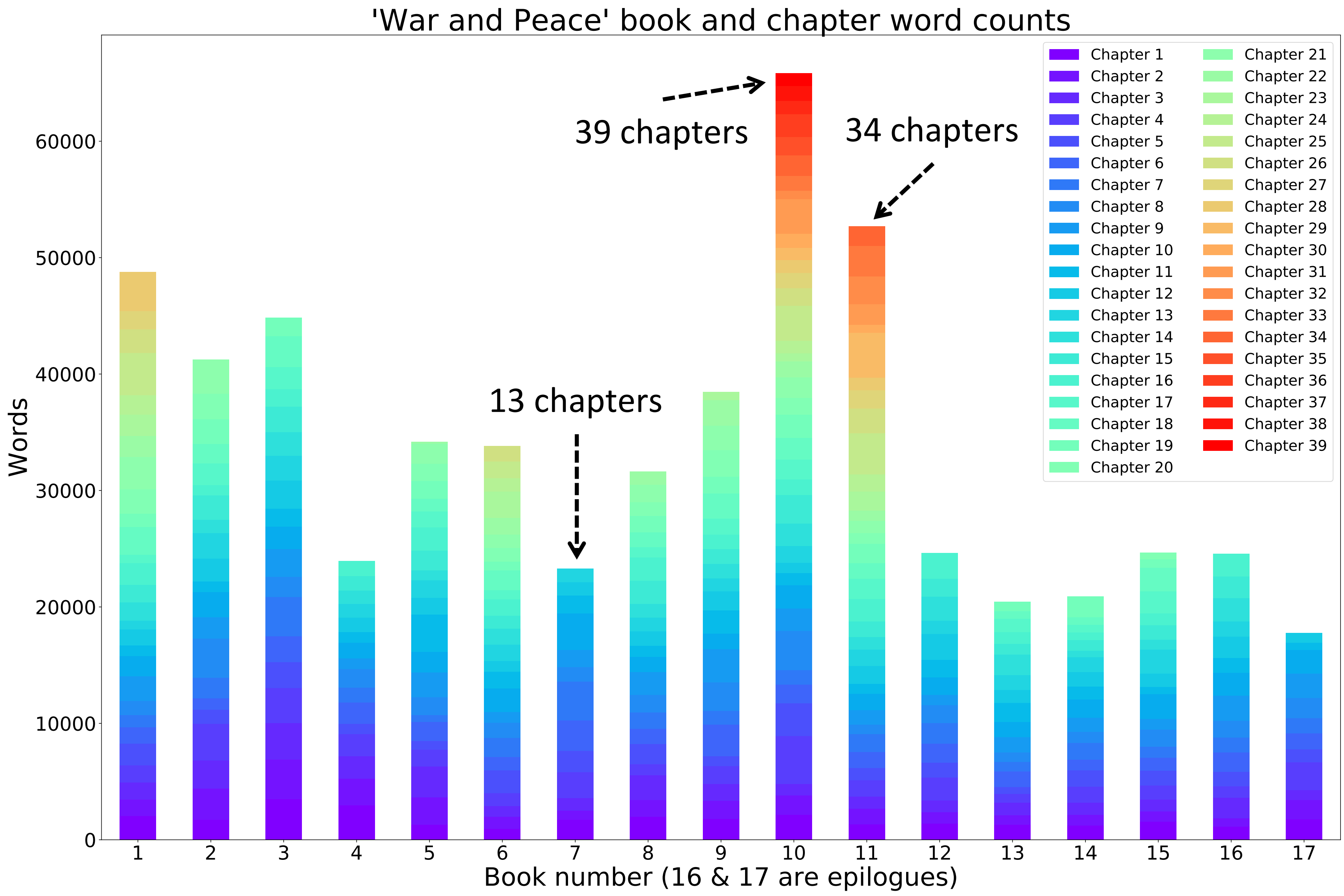
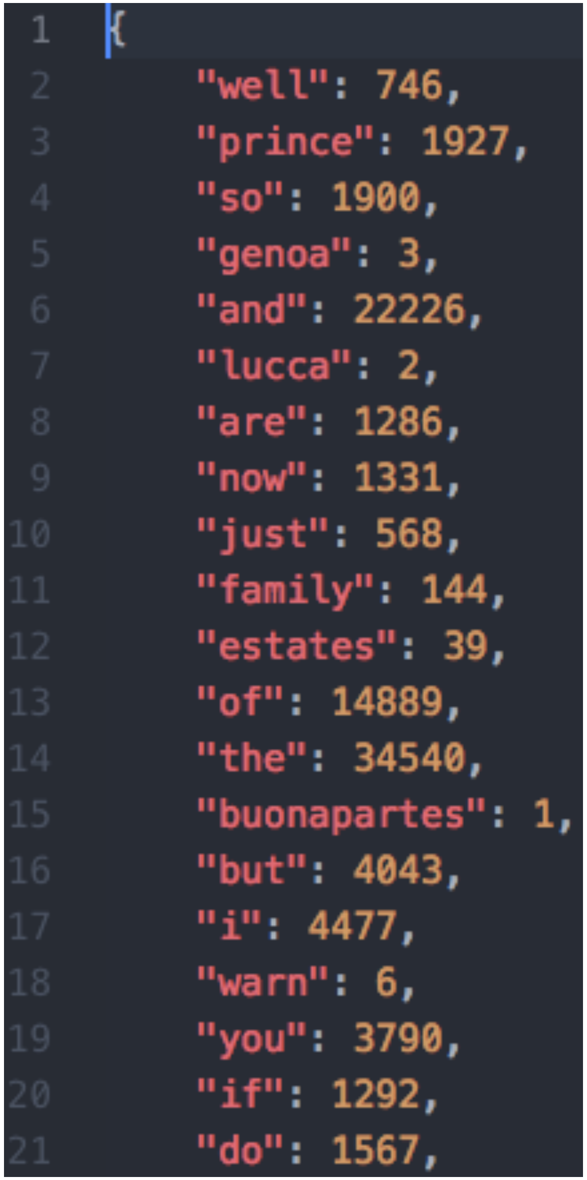
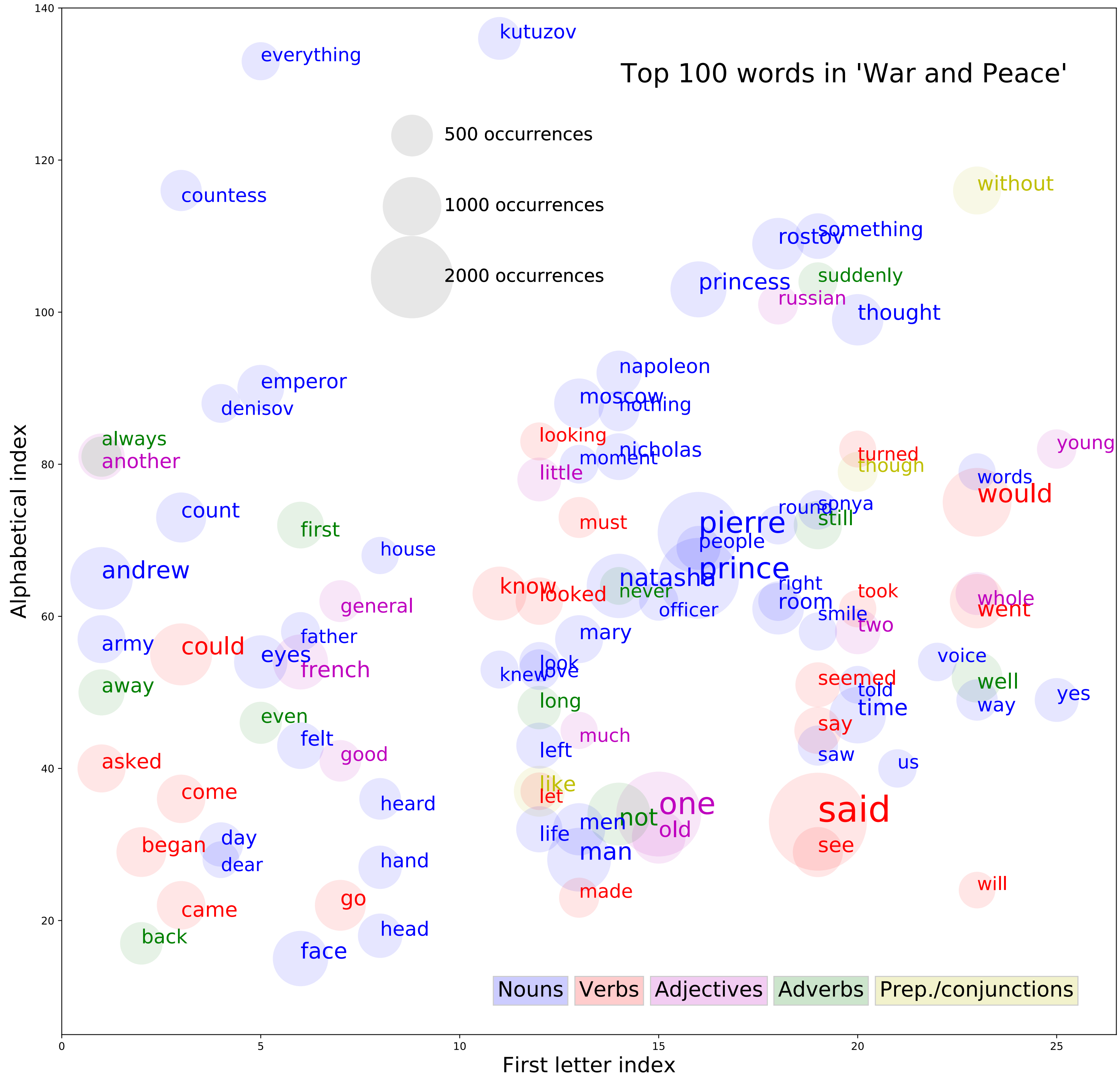


Table of word occurrences:



```
{
        "well": 746,
        "prince": 1927,
        "so": 1900,
        "genoa": 3,
        "and": 22226,
        "lucca": 2,
        "are": 1286,
        "now": 1331,
        "just": 568,
        "family": 144,
        "estates": 39,
        "of": 14889,
        "the": 34540,
        "buonapartes": 1,
        "but": 4043,
        "i": 4477,
        "warn": 6,
        "you": 3790,
        "if": 1292,
        "do": 1567,
```

Full data is in
‘word_counts.json’



Word count visualization:

- Stop words removed
- Ordered alphabetically
- Colored by part of speech (a few mistags)
- Many names (e.g. Pierre, Natasha, etc.)
- No 'I' and 'J' words in top 100
- Lots of royalty: prince, princess, emperor, etc.