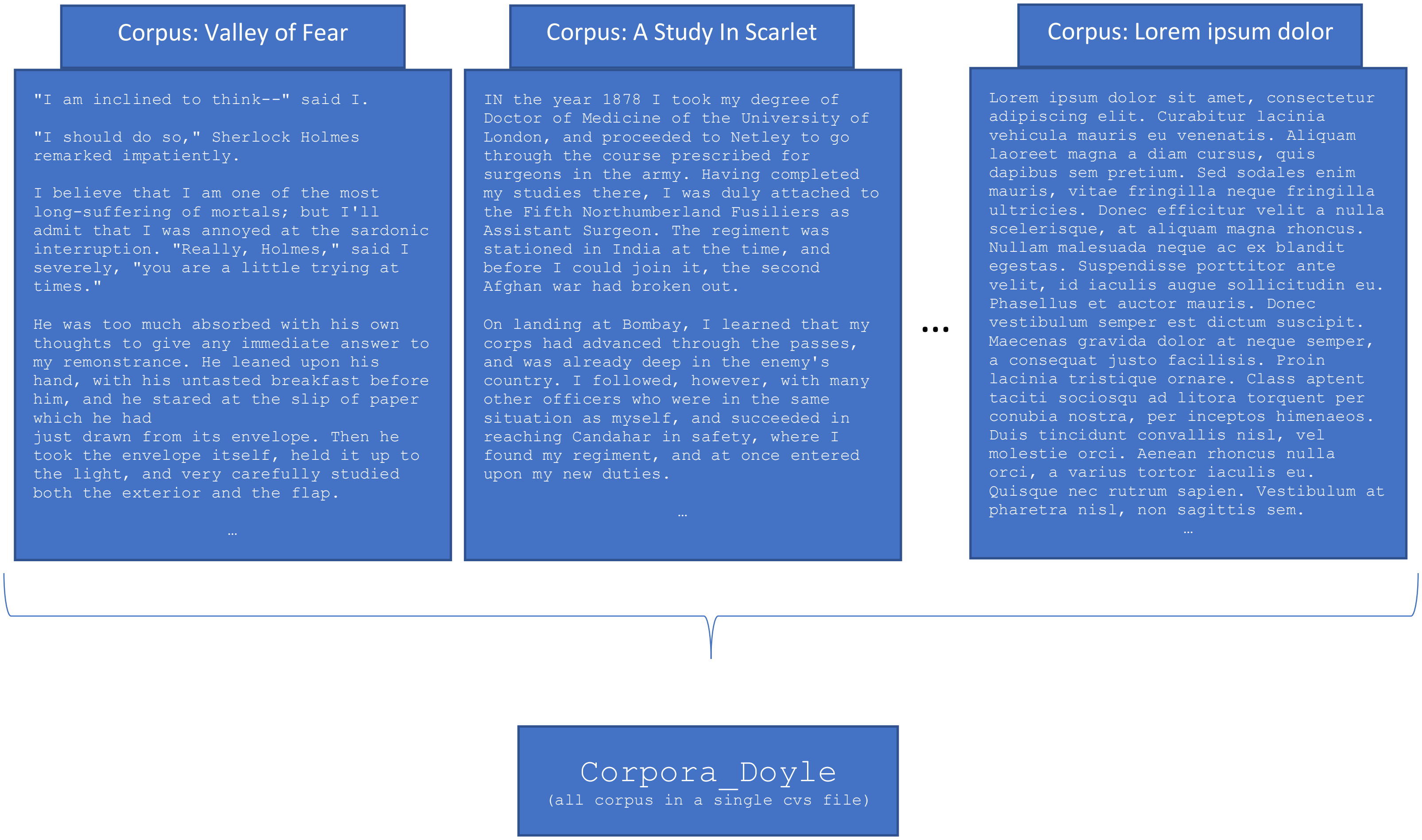
COSC 692-Nautal Language Processing
Graphical ideas – Project 2

1. We will read one file as the corpora for every author.



Corpus: Valley of Fear

"I am inclined to think--" said I.

"I should do so," Sherlock Holmes remarked impatiently.

I believe that I am one of the most long-suffering of mortals; but I'll admit that I was annoyed at the sardonic interruption. "Really, Holmes," said I severely, "you are a little trying at times."

He was too much absorbed with his own thoughts to give any immediate answer to my remonstrance. He leaned upon his hand, with his untasted breakfast before him, and he stared at the slip of paper which he had
just drawn from its envelope. Then he took the envelope itself, held it up to the light, and very carefully studied both the exterior and the flap.

…

Corpus: A Study In Scarlet

IN the year 1878 I took my degree of Doctor of Medicine of the University of London, and proceeded to Netley to go through the course prescribed for surgeons in the army. Having completed my studies there, I was duly attached to the Fifth Northumberland Fusiliers as Assistant Surgeon. The regiment was stationed in India at the time, and before I could join it, the second Afghan war had broken out.

On landing at Bombay, I learned that my corps had advanced through the passes, and was already deep in the enemy's country. I followed, however, with many other officers who were in the same situation as myself, and succeeded in reaching Candahar in safety, where I found my regiment, and at once entered upon my new duties.

…

…

Corpus: Lorem ipsum dolor

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur lacinia vehicula mauris eu venenatis. Aliquam laoreet magna a diam cursus, quis dapibus sem pretium. Sed sodales enim mauris, vitae fringilla neque fringilla ultricies. Donec efficitur velit a nulla scelerisque, at aliquam magna rhoncus. Nullam malesuada neque ac ex blandit egestas. Suspendisse porttitor ante velit, id iaculis augue sollicitudin eu. Phasellus et auctor mauris. Donec vestibulum semper est dictum suscipit. Maecenas gravida dolor at neque semper, a consequat justo facilisis. Proin lacinia tristique ornare. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Duis tincidunt convallis nisl, vel molestie orci. Aenean rhoncus nulla orci, a varius tortor iaculis eu. Quisque nec rutrum sapien. Vestibulum at pharetra nisl, non sagittis sem.

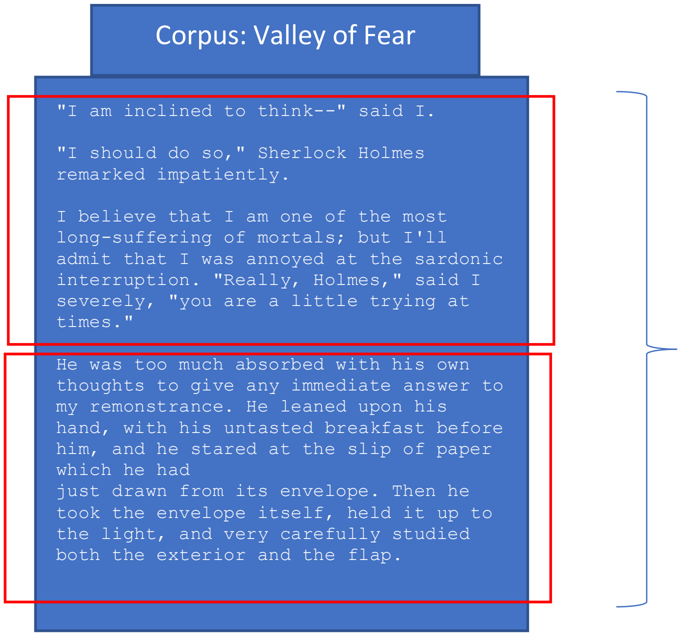
…

Corpora_Doyle
(all corpus in a single cvs file)

Technical details:

- We need that corpora as a **csv** file where every row is a sentence from the novels.
- Name the column in the csv file as “spoken_words”.
- This csv is necessary to train one model per author.

2. Documents per corpus. You will define what is a document. For example:



Corpus: Valley of Fear

"I am inclined to think--" said I.

"I should do so," Sherlock Holmes remarked impatiently.

I believe that I am one of the most long-suffering of mortals; but I'll admit that I was annoyed at the sardonic interruption. "Really, Holmes," said I severely, "you are a little trying at times."

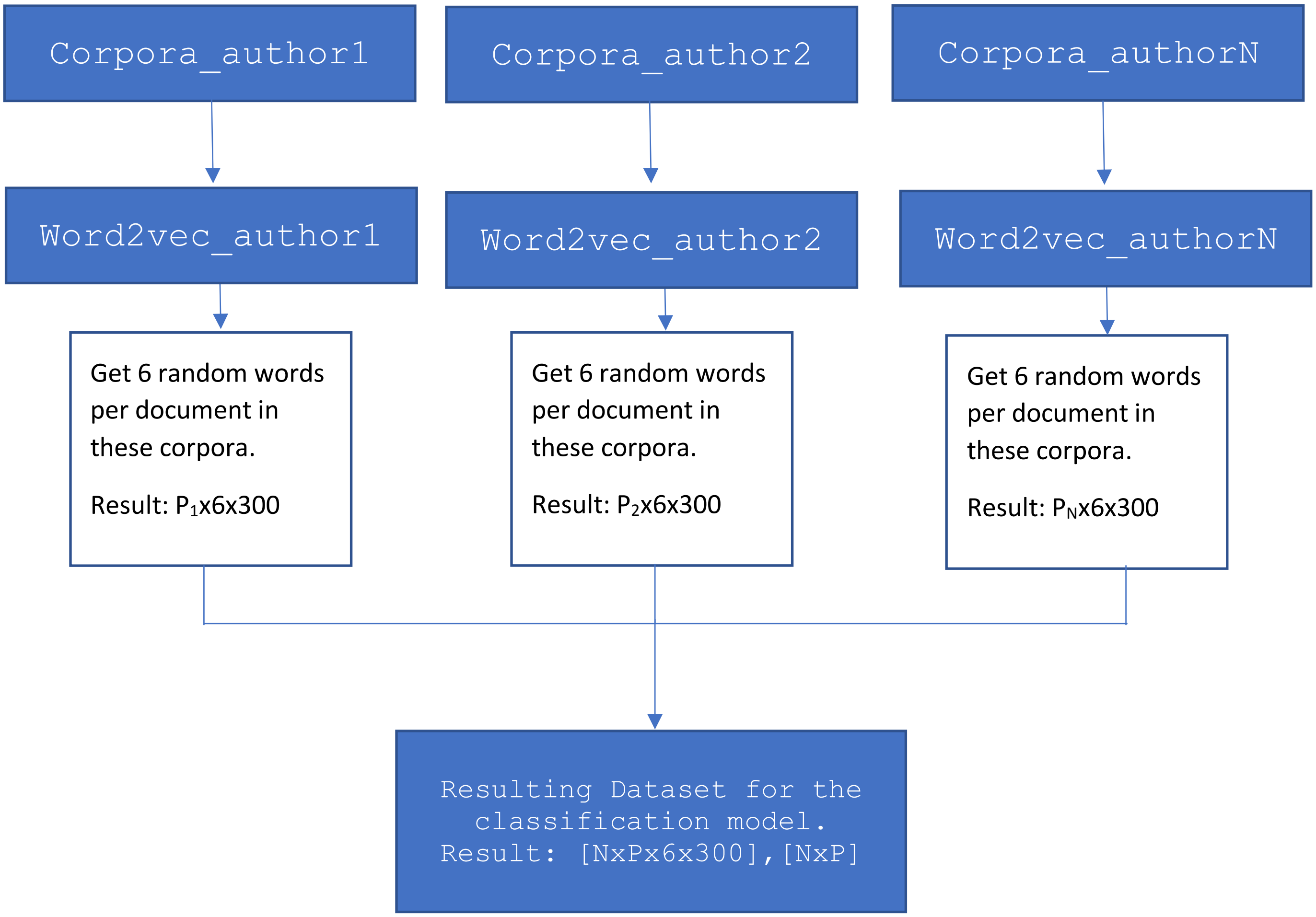
He was too much absorbed with his own thoughts to give any immediate answer to my remonstrance. He leaned upon his hand, with his untasted breakfast before him, and he stared at the slip of paper which he had
just drawn from its envelope. Then he took the envelope itself, held it up to the light, and very carefully studied both the exterior and the flap.

You can decide to say: one document is one paragraph (clearly this is not realistic), as in the case of these red boxes.

Or you can say: one document is one chapter (in my opinion this is not good because chapter are very long).

What we need from you is a method that will be able to choose ‘n’ number of words randomly from one document. This is for the final pipeline we need to build.

3. Pipeline (according to the paper).


Corpora_author1
Corpora_author2
Corpora_authorN
Word2vec_author1
Word2vec_author2
Word2vec_authorN
Get 6 random words per document in these corpora.
Result: P_1x6x300
Get 6 random words per document in these corpora.
Result: P_2x6x300
Get 6 random words per document in these corpora.
Result: P_Nx6x300
Resulting Dataset for the classification model.
Result: [NxPx6x300],[NxP]

Technical details:

- N: authors.
- P: total of documents.
- NxPx6x300: total number of samples.
- NxP: labels. Just note that the labels correspond to the different authors.