HW4 & Final Project

Date: 2019/12/9

HW4

- Here we offer the dataset, Google AI published Research.
- In this dataset, we only offer ‘title’ and ‘abstract’ and concatenate both (Google_AI_published_research.csv).
  - ${Title}. ${Abstract}
  - 2298 data samples

Word Preprocessing

- 請在report描述你試了哪些方法以及理由
- The process of transform document into word vectors and get useful information
  - Clean Stopword: 清除不影響理解的文字，例如：a, an, the, is, etc.
  - Capitalization: 將所有字轉換成小寫。
  - Stemming: 將句子分段。
  - Lemmatization: 將字義接近的字轉為原型。
    - the boy’s cars are different colors
    - the boy car be differ color

Stemming

- 如何取樣一個subword
- Example:

“Word embedding is the modern way of representing words as vectors.”

↓

["Word", "embedding", "is", "the", "modern", "way", "of", "representing", "words", "as", "vectors"]

↓

["Word embedding", "is the", "modern way", "of representing", "words as", "vectors"]

↓

["Word embedding", "is", "the", "modern", "way", "of", "re", "presenting", "words", "as", "vectors"]

↓

["Word", "embed", "modern", "way", "re", "present", "word", "vector"]

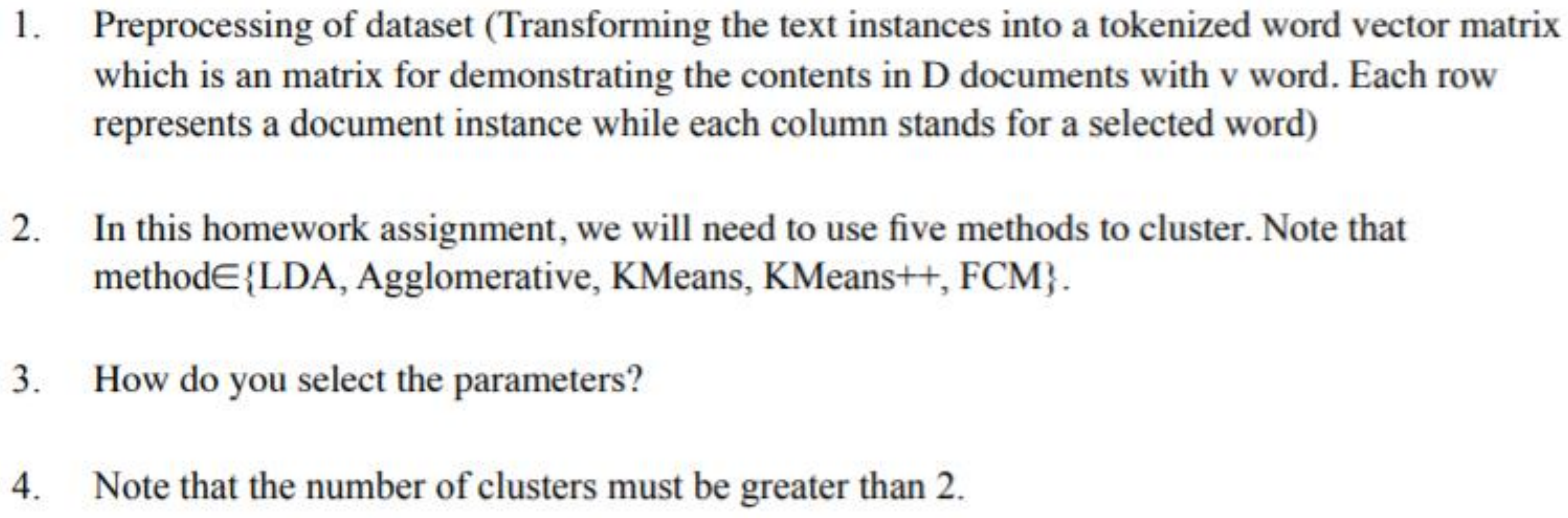
- 請在report裡描述Tokenize的過程以及理由。

Tokenizer

- 將word轉換成integer or one-hot的步驟
- Example:
  - ML is interesting. → [0, 1, 2]
  - DL is interesting. → [3, 1, 2]

評分標準

- 基本分 60分
- 報告內容 5~10分
- 剩餘30分依作業說明細項配分



1. Preprocessing of dataset (Transforming the text instances into a tokenized word vector matrix which is an matrix for demonstrating the contents in D documents with v word. Each row represents a document instance while each column stands for a selected word)

2. In this homework assignment, we will need to use five methods to cluster. Note that method∈{LDA, Agglomerative, KMeans, KMeans++, FCM}.

3. How do you select the parameters?

4. Note that the number of clusters must be greater than 2.

Final Project

- Proposal
  - 一組只需一人代表上傳
- Dataset
  - Project使用的dataset
- Code & report
  1. All in ipynb
  2. Code in ipynb & report in PDF
- Oral
  - 5~10 mins 短片 介紹project內容

Final Project

- Proposal
  - 動機, 想解決的問題是?
  - Dataset
  - Method & model
  - 預期的結果

- 假如是與專題相關的題目，禁止完全相同，但是可以從相同題目延伸
  - Ex. 專題針對情境A做到performance的提升，進一步簡化模型同時維持接近的performance