Math 444: Homework 6

Iris Zhang

April 15, 2020

Problem 1

We define a list of stop words that will be eliminated from the documents as following:
{of, and, are, with, in, the, due, to, there, that, because, is, from, can, be, but, if, one, could, back, where, prior, again, only, than, often, after, each, some, as, a, during, for, throughout, an, up, on, about, down, However, all, at, does, not, has, many, people, you, first, they, our, same, more}

Problem 2

Now we manually form a dictionary of words (with stemming) extracted from the union of all documents as following:

```
Dictionary = {'accumulations','snow','ice','common','associated','winter','north','hemisphere',
'large','land','masses','believe','temperature','change', 'Earth','close','sun',
'summer','far','fact','July','January','Canada','brutal','like','hokey','right','spot',
'end','approach','see','flock','geese','begin','year','migrat','spend','warm',
'months','fly','south','Cleveland','familiar','sight','April','October,'birds',
'650','species','American','breeding','half','cover','thousands','miles','annual','travel',
'course','little','deviation','spring','approximately','500,000','Sandhill','Cranes','endangered',
'Whooping','use','central','platte','river','valley','Nebraska','staging','habitat','nest',
'grounds','Alaska','Siberian','Arctic','millions','monarch','butterfly','United','States',
'California','Mexico','fall','adults','east','population','3,000','following','leave','overwinter',
'sites','lay','egg','milkweeds','way','hunkered','Wendy','park','metropark','Lakefront',
'reservation','longest','lifespan','Ohio','live','10','hundreds','vast','majority','September',
'voluntarily','economic','family','study','reasons'};
```

Problem 3

Here we manually form the term-document matrix A_{120X12} in EXCEL since we have 120 words in the above dictionary and also we are given 12 documents. The entries in the matrix A is the number of occurrence of each word in each document, i.e. A_{ij} = the number of occurrence of $word_i$ in D_j where $1 \leq i \leq 120$, $1 \leq j \leq 12$.

Problem 4

We do NMF developed in Homework 5 on the matrix A that we have above using k=3,4.
For k=3, we have the first five important words for each feature vector:
{migrat, travel, birds, year, Cleveland}, {winter, Canada, migrat, fall, monarch},{sun, migrat, Canada, far, geese}.
Also, the most significant feature vector for each document is arranged as: {2,3,2,3,1,1,3,2,1,1,2,3}

For k=4, we have the first five important words for each feature vector:
{sun, migrat, months, year, winter}, {United, States, spring, north, south}, {birds, migrat, Cleveland, thousands, geese}, {year, Canada, travel, Cranes, migrat}
Also, the most significant feature vector for each document is arranged as: {2,1,3,1,3,4,4,2,2,3,1,1}

From above significant words that we have, we can roughly estimate that the documents are mainly about birds and geese migrate from country to country due to the season change.