Assignment 3: Unsupervised NLP

Team contribution:

Team members	Anh Ha	Dat Nguyen	Phuong Nguyen
Solving problems, coding	40%	60%	40%
Analysing results, reports	60%	40%	60%

1. Topic analysis:

a) Tf-idf feature extraction:

Dataset properties and parameters settings:

Dataset was run from the file named 'awards_2002' with 9923 documents.

The number of features got 53816 when TfidfVectorizer was run with default settings for all parameters.

We tested by changing ngram in range of (1,1), (2,2), (3,3), (1,3) with fixing the other parameters such as min_df=2, max_df=1.0, stop_words = 'english', lowercase= True,use_idf=True, sublinear_tf=True.

We changed min_df = 5 and ngram_range = (1,3)

The evaluation results:

The number of features extracted were 27001, 190731, 117124, 334856, for the case of ngram in range of (1,1), (2,2), (3,3), (1,3) respectively.

The number of features extracted was reduced from 334 856 to 54 049 when changed min_df from 2 to 5 with ngram_range = (3,3).

By investigating the key words based on tf-idf weights, the most useful keywords of phrases can be extracted by using tf-idf with ngram = (1,3) and min_df =5.

The most useful keywords of phrases for example 5 documents:

Document 0, key words by TF-IDF

0.27	chow
0.18	hodge
0.18	algebraic
0.18	algebraic geometry
0.14	geometry algebraic

Document 1, key words by TF-IDF

0.27	control engineering
0.17	diverse group students
0.17	particular students
0.17	students seeking
0.16	cultivating

Document 2, key words by TF-IDF

0.22	updating
0.18	model based simulation
0.16	based simulation
0.14	reference
0.14	proposed effort

Document 3, key words by TF-IDF

0.20	group theory
0.17	open problems
0.16	conference
0.14	computations
0.13	learn

Document 4, key words by TF-IDF

0.21	design alternative
0.18	engineering design
0.17	product line
0.16	commonality
0.15	uncertainties

b) k-means clustering:

At first, we run KMeans with default settings for parameters, *n_clusters=8, init ='k-means++', n_init=10, max_iterint=300*. We got 8 clusters for 9923 documents from the TF-idf result with ngram = (1,3) and min_df=5. We tested with the various numbers of clusters like 10, 15, 20.

We can observe topics that are stable across multiple runs:

1. *nmr, sensor, spin, polymer, quantum, wireless, agents, nano, reactions, routing*
2. *fellowship, sciences fellowship, mathematical sciences fellowship, mathematical sciences, mathematical, sciences, postdoctoral fellowship, biological informatics, postdoctoral, informatics*
3. *manifolds, equations, algebras, algebraic, hyperbolic, spaces, number theory, representation theory, conjecture, string*
4. *available, zygotic, zygomycota aftol project, zygomycota aftol, zygomycota, zygomycetes ascomycetes basidiomycetes, zygomycetes ascomycetes, zygomycetes, zygmund operators spaces, zygmund operators*
5. *ice, contract, galaxies, mantle, stars, arctic, solar, magma, galaxy, ocean*
6. *fuel, phase ii project, sensor, drug, phase ii, ii project, coatings, fuel cell, silicon, drug delivery*
7. *genes, species, plants, brain, plant, protein, genome, political, genetic, arabidopsis*
8. *workshop, conference, stem, reu, teachers, symposium, meeting, reu site, mathematics, scholarship*

Therefore, the number of clusters gives the best picture of the topics in the corpus is 8.

c) Based on the results of a) and b), list the 10 first clusters and keywords:

Cluster: 0 (751 docs)
workshop, conference, symposium, meeting, french, cnrs, gordon, congress, igert, speakers
Cluster: 1 (1628 docs)
wireless, grid, software, sensor, power, code, visual, mobile, programming, visualization
Cluster: 2 (1681 docs)
galaxies, spin, stars, magnetic, contract, nmr, polymer, molecules, reactions, galaxy
Cluster: 3 (1075 docs)
stem, reu, teachers, reu site, scholarship, mathematics, teacher, csems, girls, learning
Cluster: 4 (225 docs)
available, zurich, zr, zooplankton species, zooplankton, zoology, zoological, zoning, zones long lived, zones long
Cluster: 5 (442 docs)
phase ii, fuel, phase ii project, ii project, sensor, coatings, membrane, drug, optical, fuel cell
Cluster: 6 (1474 docs)
ice, mantle, ocean, arctic, solar, wind, fault, seismic, forest, magma
Cluster: 7 (896 docs)
political, children, social, firms, archaeological, language, decision, policy, labor, organizational
Cluster: 8 (597 docs)
sciences fellowship, mathematical sciences fellowship, mathematical sciences, fellowship, manifolds, mathematical, equations, sciences, spaces, algebraic
Cluster: 9 (1154 docs)
genes, species, protein, plants, proteins, fellowship, genome, plant, genetic, gene

d) From top 10 clusters,

Two good clusters:

1.wireless, grid, software, sensor, power, code, visual, mobile, programming, visualization
2.ice, mantle, ocean, arctic, solar, wind, fault, seismic, forest, magma
---> Because they have high scores mean the topics occurred in many different documents. They can illustrate the main topic related to one type of sensor applied in natural areas without reading all documents.

Two bad clusters:

1.available, zygotic, zygomycota aftol project, zygomycota aftol, zygomycota, zygomycetes ascomycetes basidiomycetes, zygomycetes ascomycetes, zygomycetes, zygmund operators spaces, zygmund operators
2.fuel, phase ii project, sensor, phase ii, ii project, drug, coatings, fuel cell, silicon, membrane
---> Because they have low scores mean the topics occurred in a few documents. They can misunderstand the main topics.

2. Word vectors:

a) 5 arbitrary words are:

-Train word2vec vectors on the corpus with parameters sg= 0,1; vector_size = 10, 100; min_count= 1, 5.
-We evaluated the results based on the higher cosine similarity and coherence of similar words with the chosen seed words.

Results & Conclusions:

1.The lists of seed words are shown differently depending mostly on min_count settings, and whatever the settings of size or sg.
For example:
min_count=5, Seed_words: ['greatly', 'every', 'includes', 'transcription', 'largest']
min_count=1, Seed words: ['half', 'whether', 'neuroscience', 'Lonza', 'tightly']

2.Size = 10 gives better results than size = 100
For example:
size=10, min_count=1, sg=1:
Most similar to: **neuroscience**
[('furthering', 0.9842066764831543), ('informational', 0.9749946594238281), ('relevance', 0.9741153717041016), ('neurobiology', 0.96800696849823), ('interest', 0.9663718342781067), ('informatics', 0.9641363620758057), ('technological', 0.9639126062393188), ('advancing', 0.9617303609848022), ('playing', 0.9597301483154297), ('sciences', 0.9593894481658936)]

size=100, min_count=1, sg=1:

Most similar to: neuroscience

[('endocrinology', 0.8928981423377991), ('neurobiology', 0.8739007115364075), ('neuroendocrinology', 0.8608860373497009), ('sociology', 0.8580520749092102), ('linguistics', 0.8549712896347046), ('epidemiology', 0.8538591861724854), ('neurophysiology', 0.8537481427192688), ('revolutionizing', 0.8456230163574219), ('archeology', 0.8387148976325989), ('informs', 0.8276082277297974)]

3.In comarison with min_count =5, min_count=1 returned more high cosine similarity result cases.

4.sg =1 returned higher cosine similarity than sg = 0

For example:

size=10, min_count=1, sg=0:

Most similar to: **neuroscience**

[('populace', 0.9711309671401978), ('LCLUC', 0.9577138423919678), ('focusing', 0.9521087408065796), ('0231010', 0.9496864676475525), ('macroeconomics', 0.9477670192718506), ('lifeline', 0.9440898895263672), ('codimension', 0.9428080320358276), ('sustainability', 0.941412627696991), ('naturalization', 0.9316829442977905), ('arena', 0.9310978651046753)]

In conclusion, for the dataset 9923 documents, the best parameters for word2vec: size= 10, min_count=1, sg=1

b) Increasing the training data:

- For the dataset named 'awards_2002' with 9923 documents, with the settings: size=10, min_count=1, sg=1, workers=4, the list of seed words: **['half', 'whether', 'neuroscience', 'Lonza', 'tightly'],** and the size of the list of vocabulary dictionaries is 63538.

The cosine similarity results:

Most similar to: **neuroscience**

[('furthering', 0.9842066764831543), ('informational', 0.9749946594238281), ('relevance', 0.9741153717041016), ('neurobiology', 0.96800696849823), ('interest', 0.9663718342781067), ('informatics', 0.9641363620758057), ('technological', 0.9639126062393188), ('advancing', 0.9617303609848022), ('playing', 0.9597301483154297), ('sciences', 0.9593894481658936)]

- For the full dataset named 'abstracts' with 132372 documents, with the settings: size=10, min_count=1, sg=1, workers=4, the list of seed words: **['Conceptual', 'fluent', 'seventh', 'manage', 'publicity']** and the size of the list of vocabulary dictionaries is 257022.

When we used the trained vector on the large dataset ('abstract') to check the similarity of the seed word 'neuroscience' from the vector on the smaller dataset ('awards_2002), the example of the cosine similarity results is shown as below:

Most similar to: **neuroscience**

[('neurobiology', 0.9721717834472656), ('psychology', 0.971102774143219), ('interdisciplinarity', 0.9698944091796875), ('imaginative', 0.9662922620773315), ('cultivating', 0.9651345014572144), ('neurosciences', 0.9640445709228516), ('metacognition', 0.958501935005188), ('collegiality', 0.9539251327514648), ('psychobiology', 0.9532018899917603), ('biopsychology', 0.9521270990371704)]

In conclusions,

- The seed words list is different from the large and small training data
- The word2vec worked on the large training dataset returned more coherence similar word with the chosen seed word than on the small training dataset
- The same dataset and same parameters settings after being tested on different computers returned the different seed word list. Therefore, the computers ' IP address may affect the results.
  For example: the list of seed words for the full dataset named 'abstracts' with 132372 documents on another computer: **['Mysticete', 'whales', 'to', 'near', 'extinction']**