Topic Modeling & Document Ranking

Ngoc Phan

M.S. Business Analytics
nphan20181@gmail.com

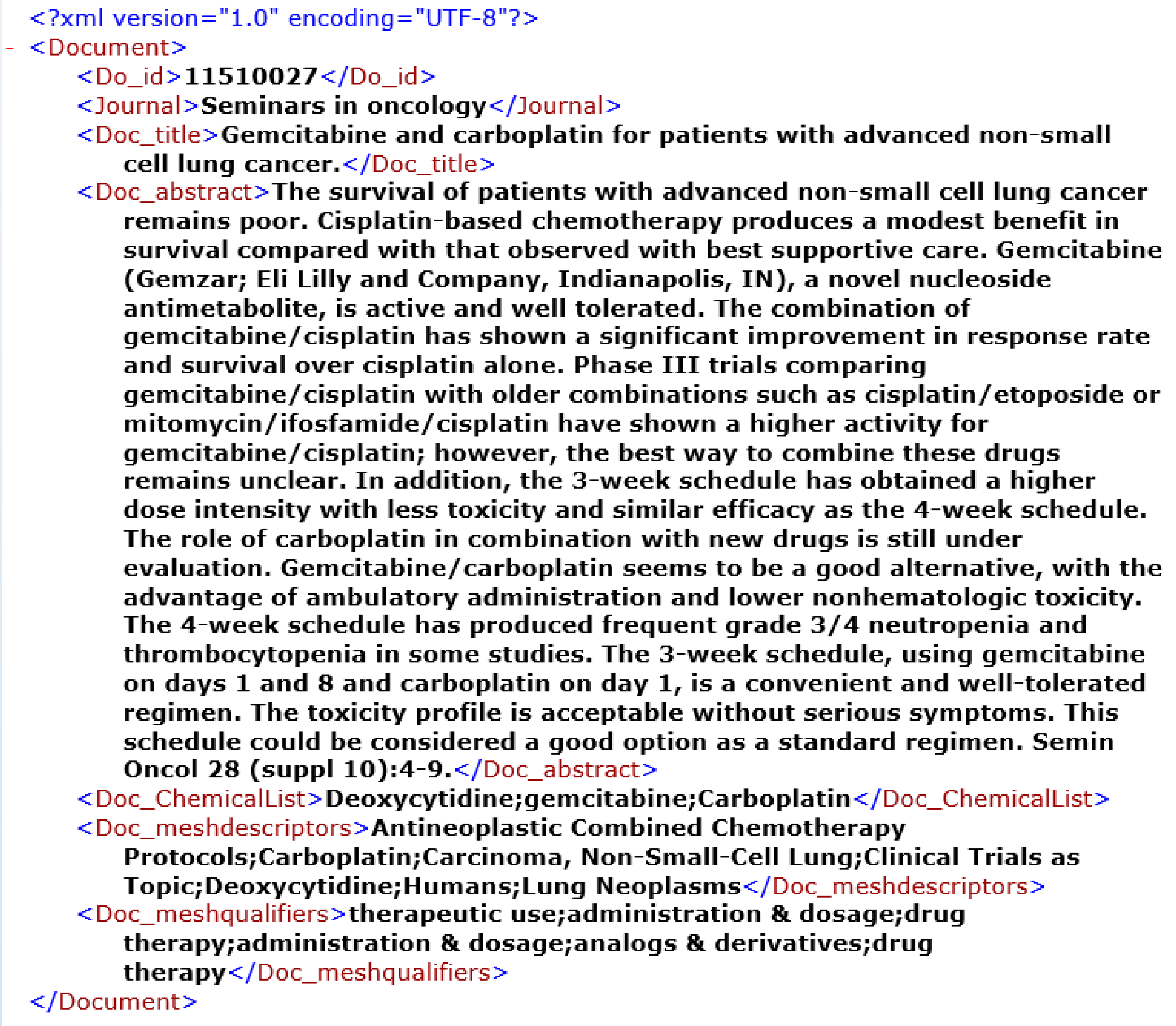
Python Source Code on GitHub
https://github.com/nphan20181/Topic_Modeling_n_Document_Ranking

PubMed
Text Retrieval Conference (TREC)

Dataset

Abstracts of Scientific Articles

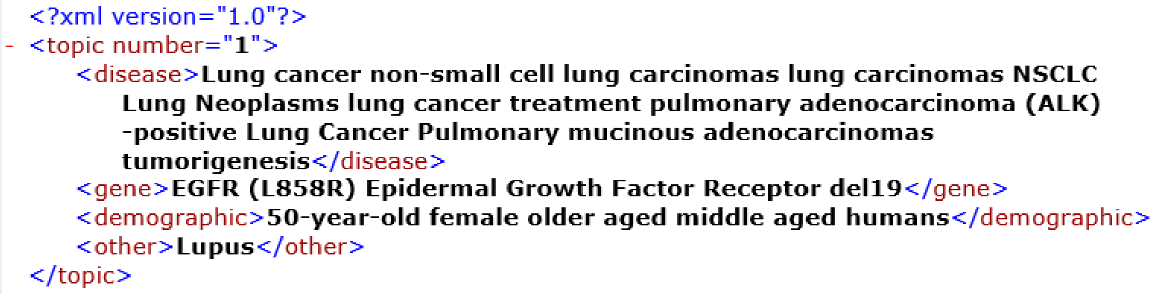
- 333 xml documents
- Information about a particular type of cancer



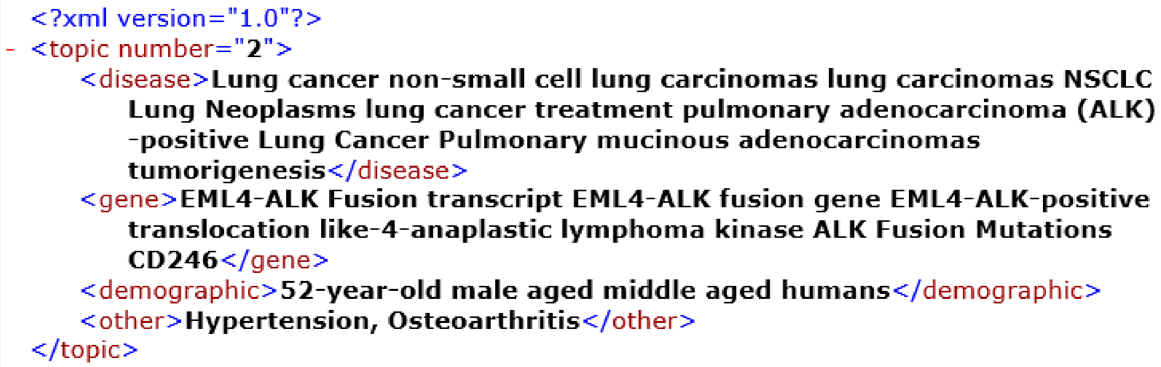
```
  <?xml version="1.0" encoding="UTF-8"?>
- <Document>
    <Do_id>11510027</Do_id>
    <Journal>Seminars in oncology</Journal>
    <Doc_title>Gemcitabine and carboplatin for patients with advanced non-small
      cell lung cancer.</Doc_title>
    <Doc_abstract>The survival of patients with advanced non-small cell lung cancer
      remains poor. Cisplatin-based chemotherapy produces a modest benefit in
      survival compared with that observed with best supportive care. Gemcitabine
      (Gemzar; Eli Lilly and Company, Indianapolis, IN), a novel nucleoside
      antimetabolite, is active and well tolerated. The combination of
      gemcitabine/cisplatin has shown a significant improvement in response rate
      and survival over cisplatin alone. Phase III trials comparing
      gemcitabine/cisplatin with older combinations such as cisplatin/etoposide or
      mitomycin/ifosfamide/cisplatin have shown a higher activity for
      gemcitabine/cisplatin; however, the best way to combine these drugs
      remains unclear. In addition, the 3-week schedule has obtained a higher
      dose intensity with less toxicity and similar efficacy as the 4-week schedule.
      The role of carboplatin in combination with new drugs is still under
      evaluation. Gemcitabine/carboplatin seems to be a good alternative, with the
      advantage of ambulatory administration and lower nonhematologic toxicity.
      The 4-week schedule has produced frequent grade 3/4 neutropenia and
      thrombocytopenia in some studies. The 3-week schedule, using gemcitabine
      on days 1 and 8 and carboplatin on day 1, is a convenient and well-tolerated
      regimen. The toxicity profile is acceptable without serious symptoms. This
      schedule could be considered a good option as a standard regimen. Semin
      Oncol 28 (suppl 10):4-9.</Doc_abstract>
    <Doc_ChemicalList>Deoxycytidine;gemcitabine;Carboplatin</Doc_ChemicalList>
    <Doc_meshdescriptors>Antineoplastic Combined Chemotherapy
      Protocols;Carboplatin;Carcinoma, Non-Small-Cell Lung;Clinical Trials as
      Topic;Deoxycytidine;Humans;Lung Neoplasms</Doc_meshdescriptors>
    <Doc_meshqualifiers>therapeutic use;administration & dosage;drug
      therapy;administration & dosage;analogs & derivatives;drug
      therapy</Doc_meshqualifiers>
  </Document>
```

Query Topics

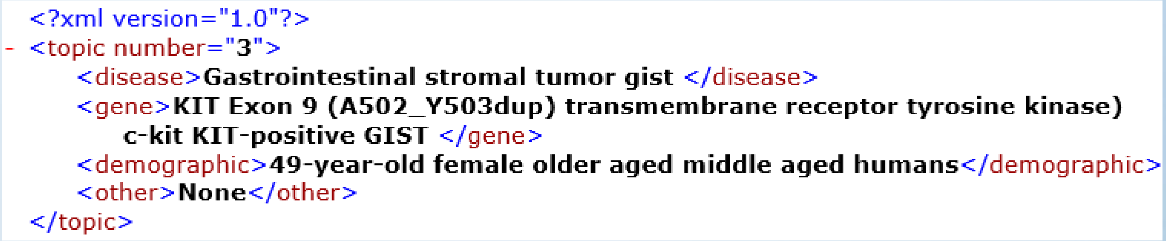
- Information about a single patient suffering from a disease
- Topic 1
  - **lung cancer, egfr, aged female**
- Topic 2
  - **lung cancer, eml4-alk, aged male**
- Topic 3
  - **gist, kit exon, aged female**



```
 <?xml version="1.0"?>
- <topic number="1">
    <disease>Lung cancer non-small cell lung carcinomas lung carcinomas NSCLC
      Lung Neoplasms lung cancer treatment pulmonary adenocarcinoma (ALK)
      -positive Lung Cancer Pulmonary mucinous adenocarcinomas
      tumorigenesis</disease>
    <gene>EGFR (L858R) Epidermal Growth Factor Receptor del19</gene>
    <demographic>50-year-old female older aged middle aged humans</demographic>
    <other>Lupus</other>
 </topic>
```



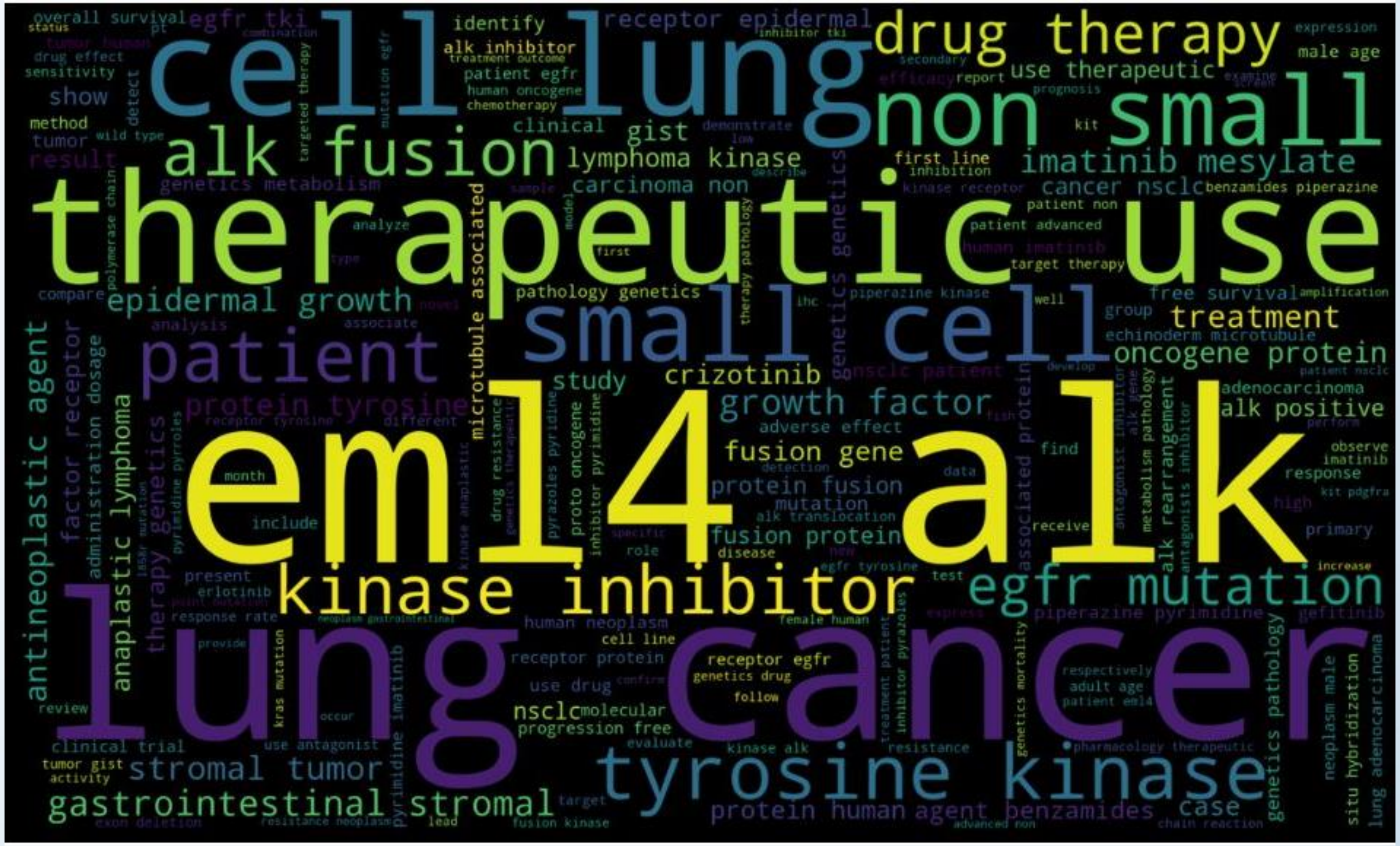
```
 <?xml version="1.0"?>
- <topic number="2">
    <disease>Lung cancer non-small cell lung carcinomas lung carcinomas NSCLC
      Lung Neoplasms lung cancer treatment pulmonary adenocarcinoma (ALK)
      -positive Lung Cancer Pulmonary mucinous adenocarcinomas
      tumorigenesis</disease>
    <gene>EML4-ALK Fusion transcript EML4-ALK fusion gene EML4-ALK-positive
      translocation like-4-anaplastic lymphoma kinase ALK Fusion Mutations
      CD246</gene>
    <demographic>52-year-old male aged middle aged humans</demographic>
    <other>Hypertension, Osteoarthritis</other>
 </topic>
```



```
 <?xml version="1.0"?>
- <topic number="3">
    <disease>Gastrointestinal stromal tumor gist </disease>
    <gene>KIT Exon 9 (A502_Y503dup) transmembrane receptor tyrosine kinase)
      c-kit KIT-positive GIST </gene>
    <demographic>49-year-old female older aged middle aged humans</demographic>
    <other>None</other>
 </topic>
```

Word Cloud

Data Analysis

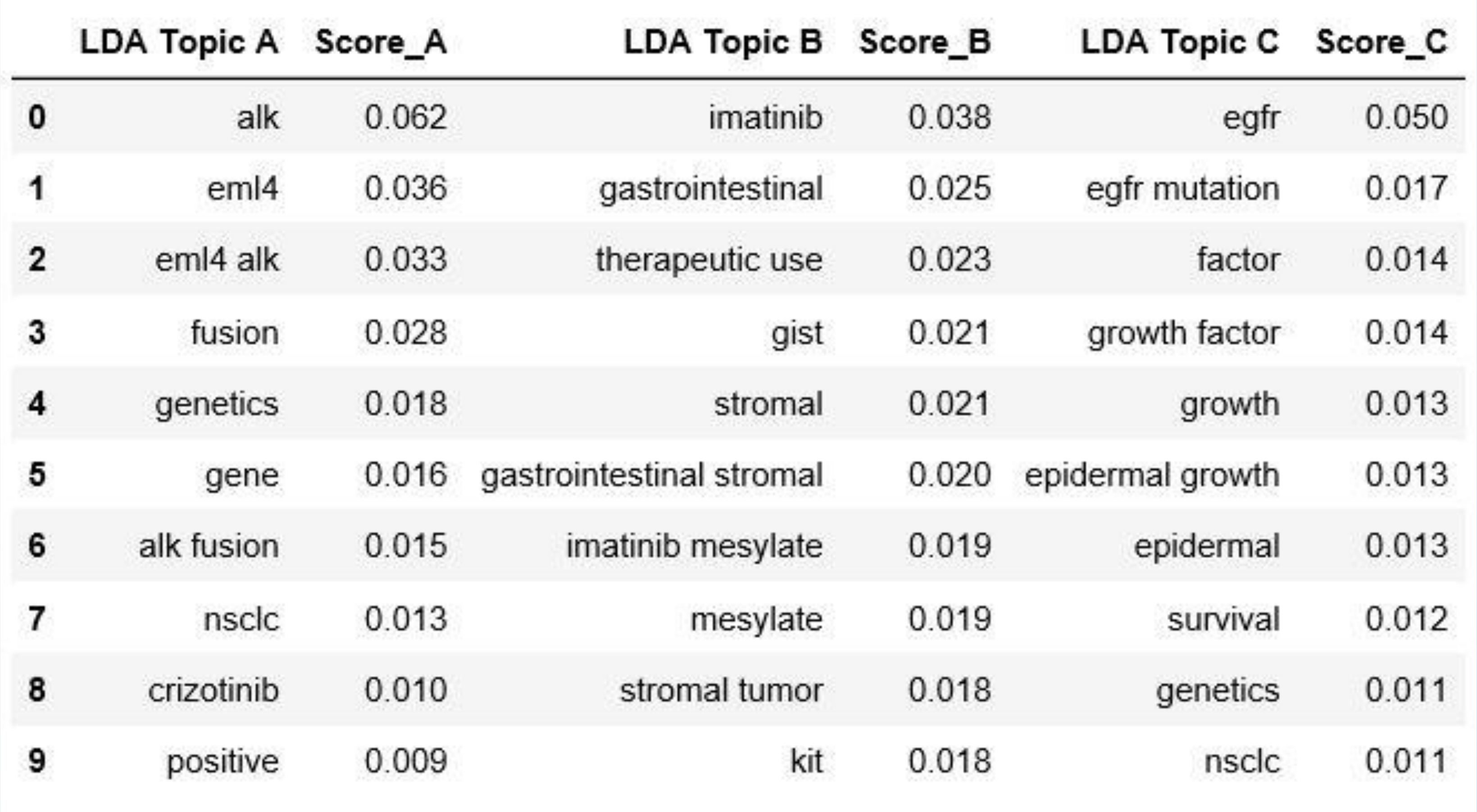


Word Cloud

Words found in the documents.
The size of each word indicates its frequency or importance.

Topic Modeling

Latent Dirichlet Allocation (LDA)



	LDA Topic A	Score_A	LDA Topic B	Score_B	LDA Topic C	Score_C
0	alk	0.062	imatinib	0.038	egfr	0.050
1	eml4	0.036	gastrointestinal	0.025	egfr mutation	0.017
2	eml4 alk	0.033	therapeutic use	0.023	factor	0.014
3	fusion	0.028	gist	0.021	growth factor	0.014
4	genetics	0.018	stromal	0.021	growth	0.013
5	gene	0.016	gastrointestinal stromal	0.020	epidermal growth	0.013
6	alk fusion	0.015	imatinib mesylate	0.019	epidermal	0.013
7	nsclc	0.013	mesylate	0.019	survival	0.012
8	crizotinib	0.010	stromal tumor	0.018	genetics	0.011
9	positive	0.009	kit	0.018	nsclc	0.011

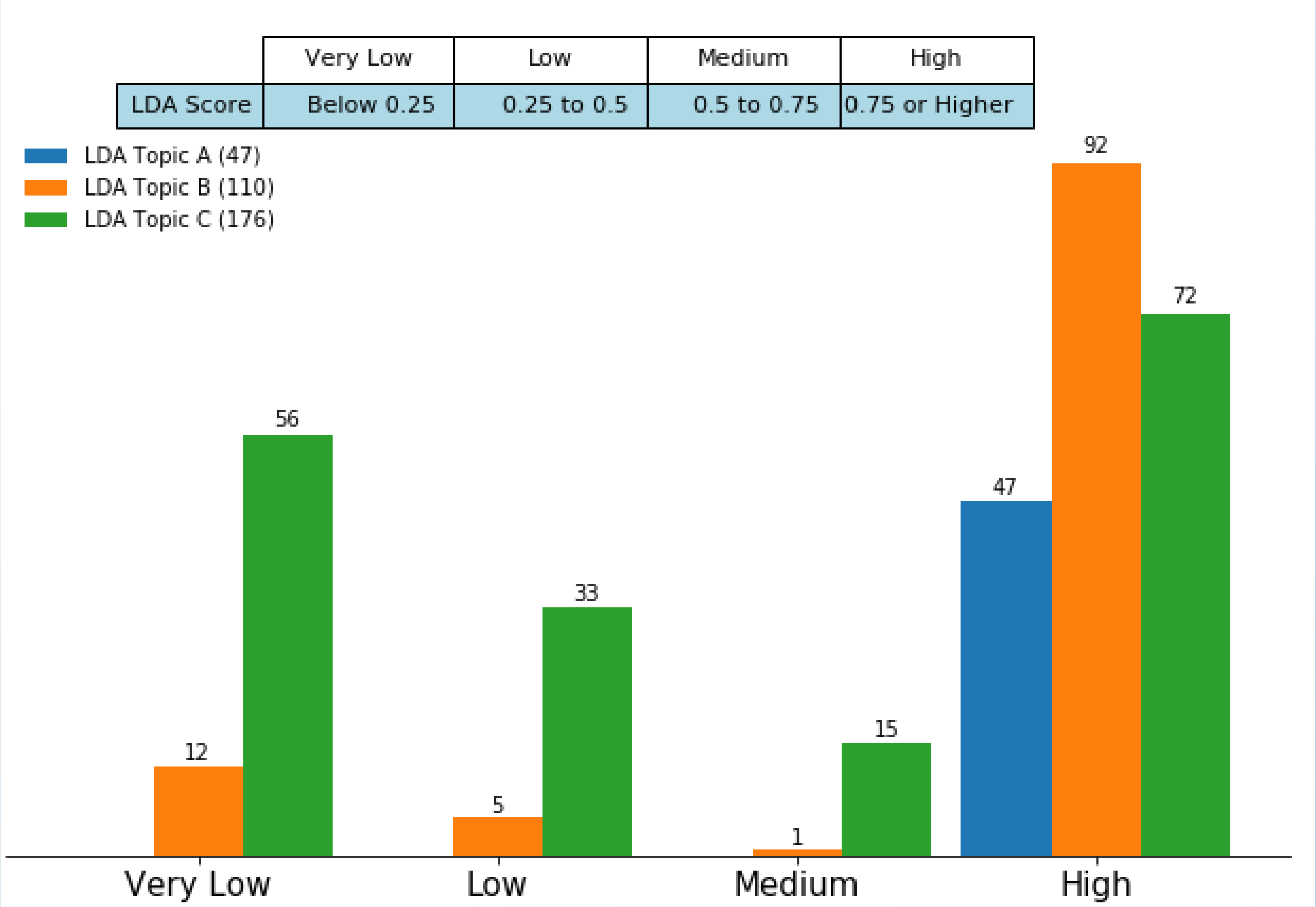
LDA Topics

- Topic A (***eml4, alk, positive***)
- Topic B (***gist, kit, therapeutic use***)
- Topic C (***egfr, epidermal, growth factor***)

Documents Distribution per LDA Score

LDA Topics

- Topic A (***eml4, alk, positive***)
- Topic B (***gist, kit, therapeutic use***)
- Topic C (***egfr, epidermal, growth factor***)



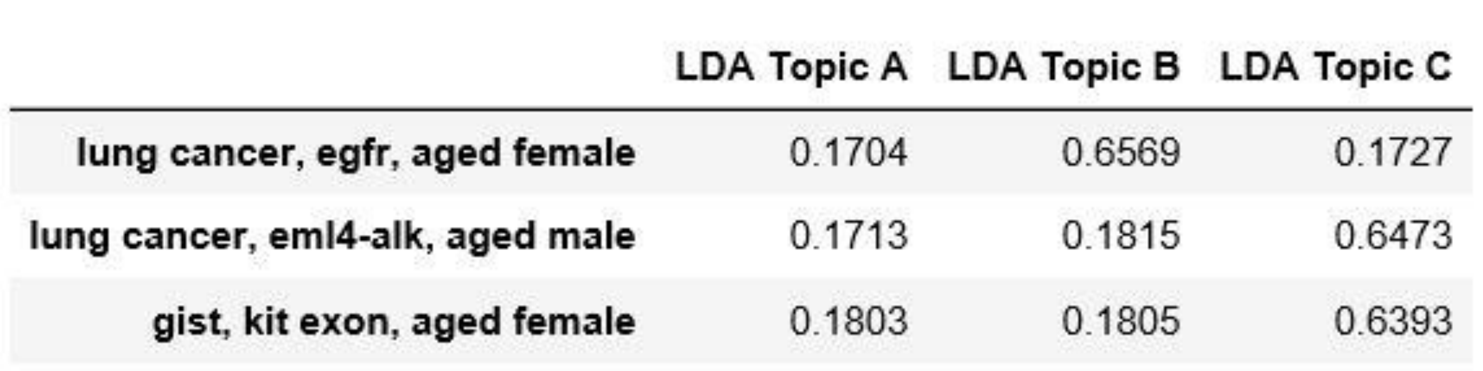
	Very Low	Low	Medium	High
LDA Score	Below 0.25	0.25 to 0.5	0.5 to 0.75	0.75 or Higher

Topic Classification

LDA Topics

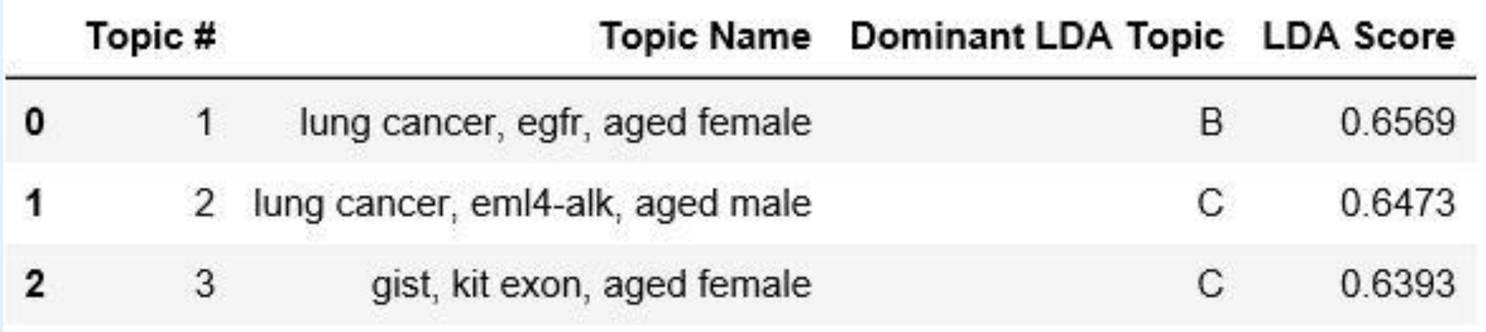
- Topic A (***eml4, alk, positive***)
- Topic B (***gist, kit, therapeutic use***)
- Topic C (***egfr, epidermal, growth factor***)

LDA Score

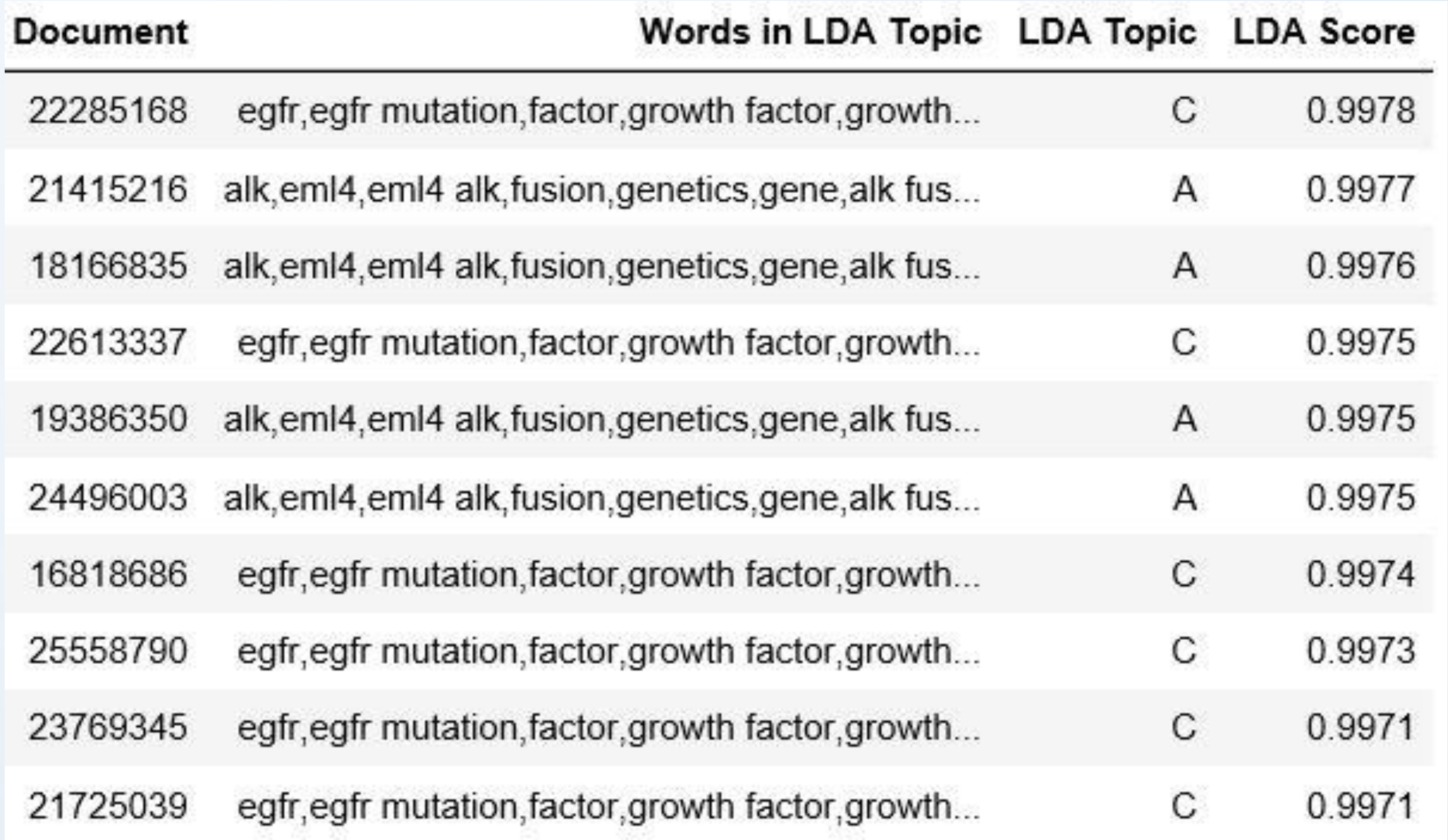


	LDA Topic A	LDA Topic B	LDA Topic C
lung cancer, egfr, aged female	0.1704	0.6569	0.1727
lung cancer, eml4-alk, aged male	0.1713	0.1815	0.6473
gist, kit exon, aged female	0.1803	0.1805	0.6393

Dominant LDA Topic



	Topic #	Topic Name	Dominant LDA Topic	LDA Score
0	1	lung cancer, egfr, aged female	B	0.6569
1	2	lung cancer, eml4-alk, aged male	C	0.6473
2	3	gist, kit exon, aged female	C	0.6393

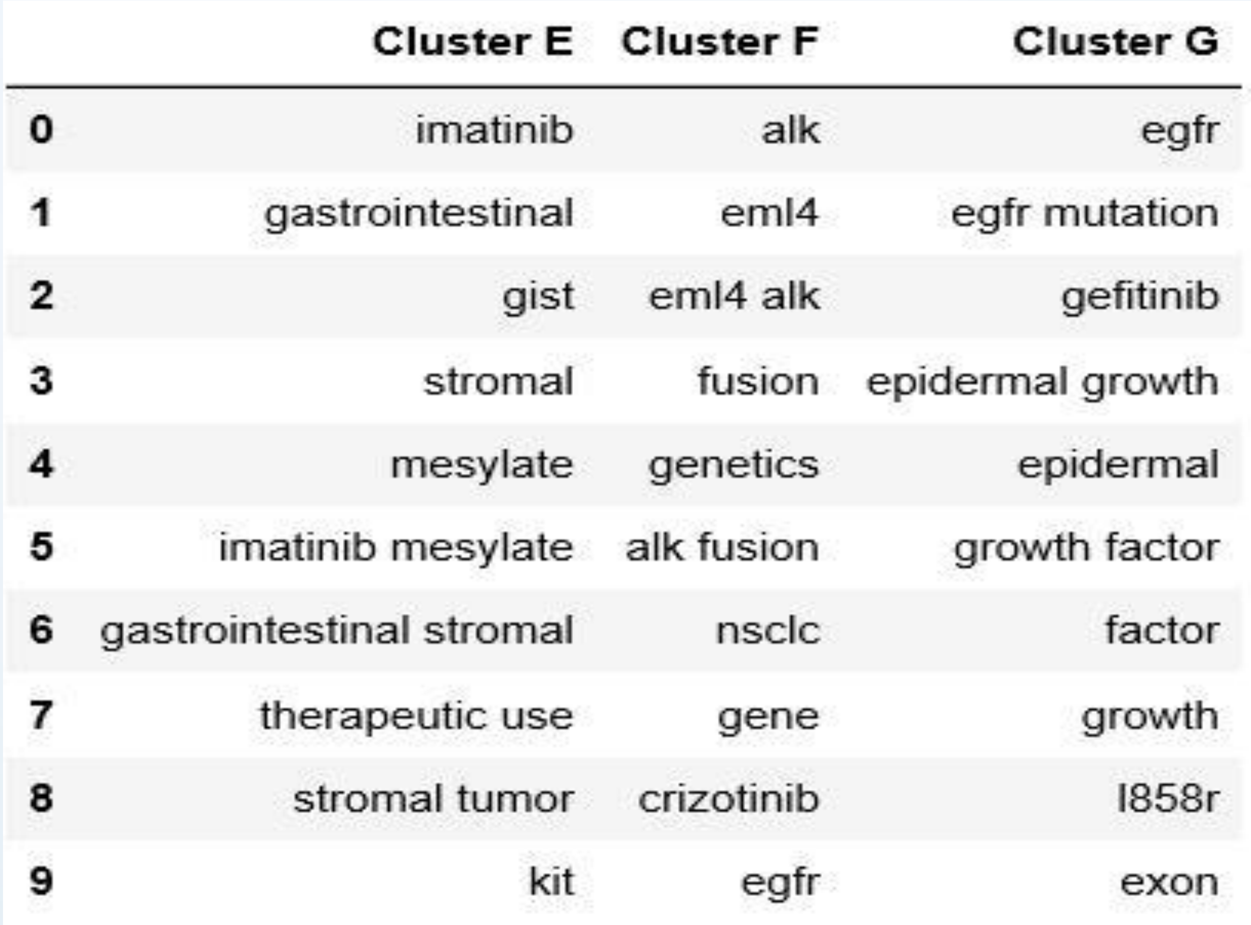


Document	Words in LDA Topic	LDA Topic	LDA Score
22285168	egfr,egfr mutation,factor,growth factor,growth...	C	0.9978
21415216	alk,eml4,eml4 alk,fusion,genetics,gene,alk fus...	A	0.9977
18166835	alk,eml4,eml4 alk,fusion,genetics,gene,alk fus...	A	0.9976
22613337	egfr,egfr mutation,factor,growth factor,growth...	C	0.9975
19386350	alk,eml4,eml4 alk,fusion,genetics,gene,alk fus...	A	0.9975
24496003	alk,eml4,eml4 alk,fusion,genetics,gene,alk fus...	A	0.9975
16818686	egfr,egfr mutation,factor,growth factor,growth...	C	0.9974
25558790	egfr,egfr mutation,factor,growth factor,growth...	C	0.9973
23769345	egfr,egfr mutation,factor,growth factor,growth...	C	0.9971
21725039	egfr,egfr mutation,factor,growth factor,growth...	C	0.9971

Documents Ranking on LDA Score

Topic Modeling

Text Clustering



	Cluster E	Cluster F	Cluster G
0	imatinib	alk	egfr
1	gastrointestinal	eml4	egfr mutation
2	gist	eml4 alk	gefitinib
3	stromal	fusion	epidermal growth
4	mesylate	genetics	epidermal
5	imatinib mesylate	alk fusion	growth factor
6	gastrointestinal stromal	nsclc	factor
7	therapeutic use	gene	growth
8	stromal tumor	crizotinib	l858r
9	kit	egfr	exon

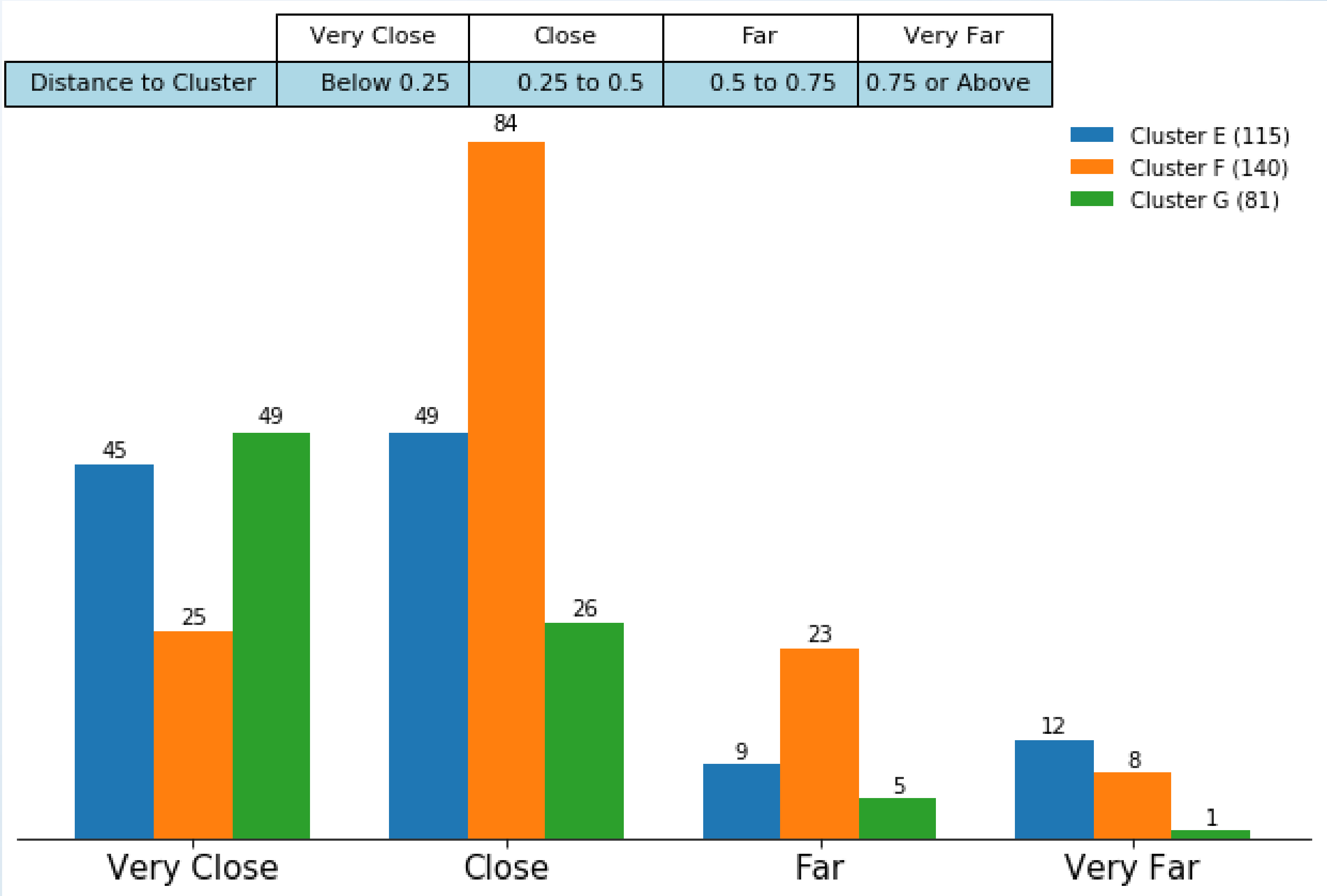
Clusters

- Cluster E (***gist, therapeutic use***)
- Cluster F (***eml4, alk, fusion, egfr***)
- Cluster G (***egfr, mutation, epidermal growth***)

Documents Distribution per Distance to Cluster

Clusters

- Cluster E (***gist, therapeutic use***)
- Cluster F (***eml4, alk, fusion, egfr***)
- Cluster G (***egfr, mutation, epidermal growth***)

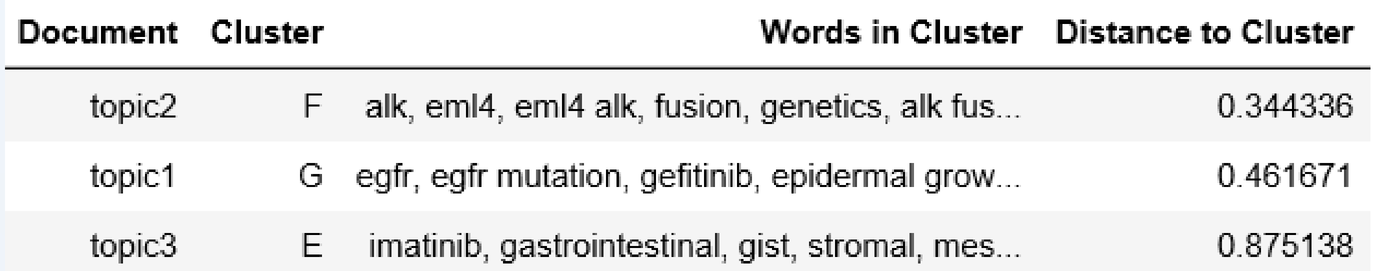


	Very Close	Close	Far	Very Far
Distance to Cluster	Below 0.25	0.25 to 0.5	0.5 to 0.75	0.75 or Above

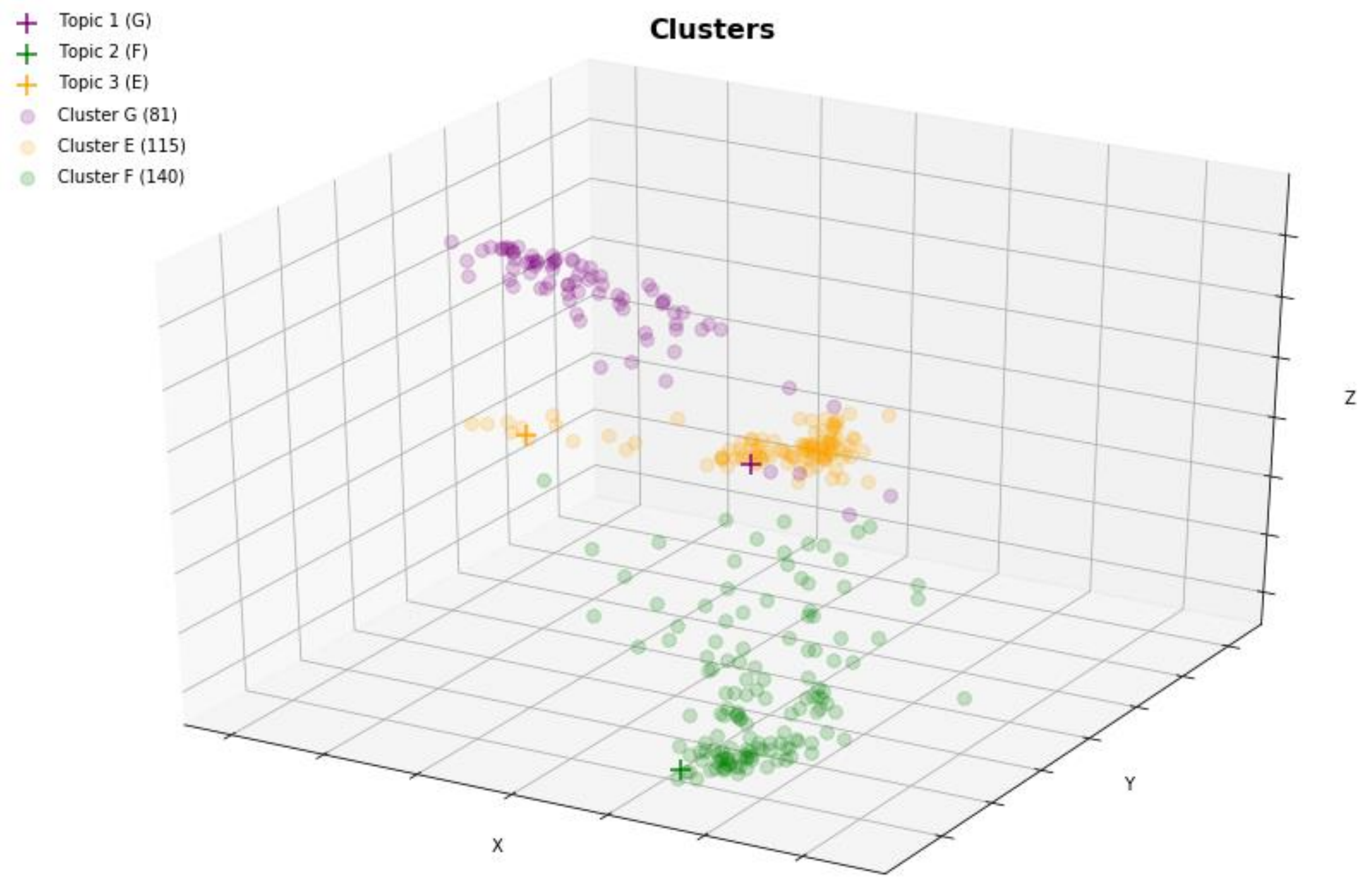
* ***Distance to Cluster*** is the distance between the document and the closest cluster.

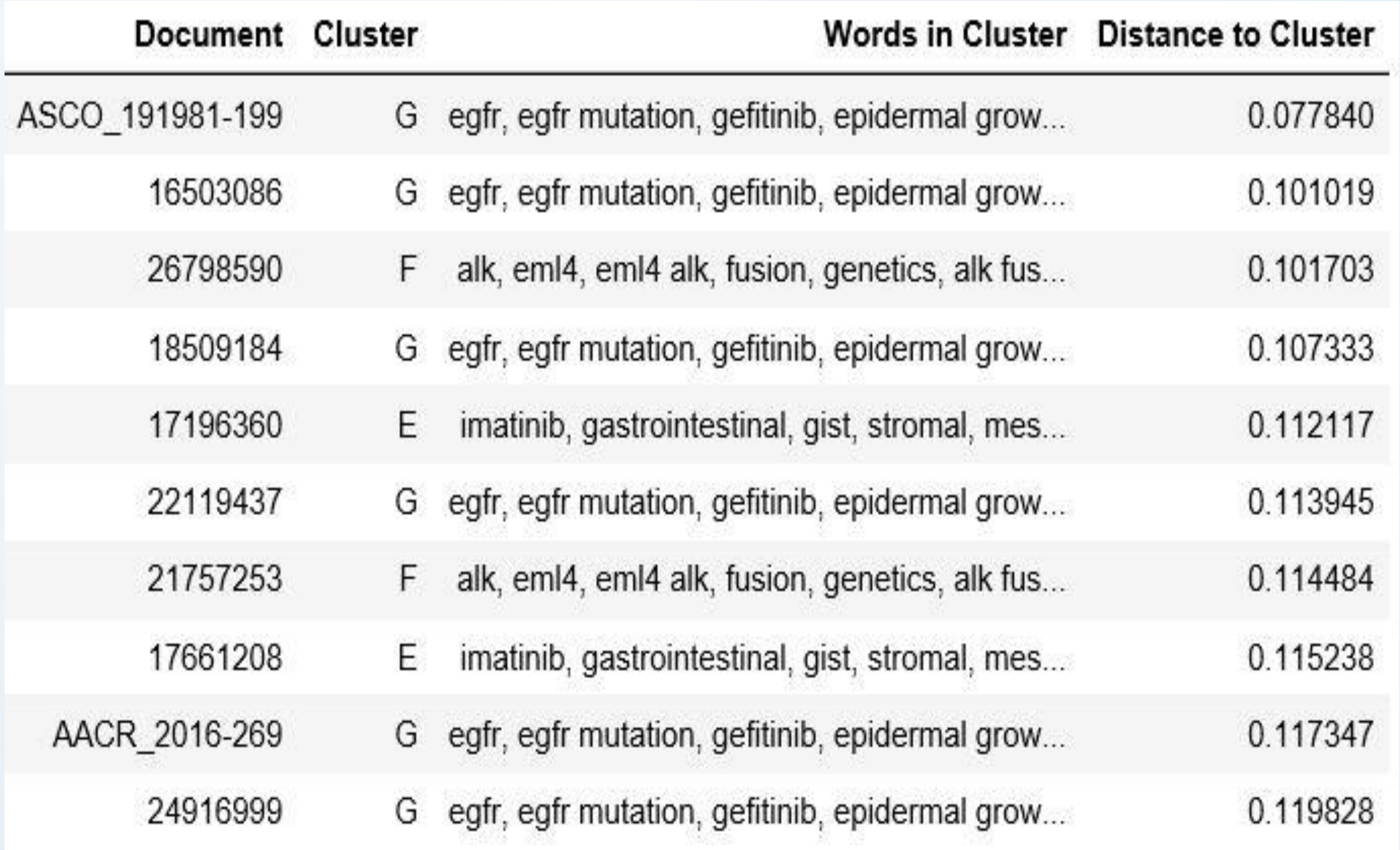
Topic Classification

- **Topic 1**: lung cancer, egfr, aged female
- **Topic 2**: lung cancer, eml4-alk, aged male
- **Topic 3**: gist, kit exon, aged female



Document	Cluster	Words in Cluster	Distance to Cluster
topic2	F	alk, eml4, eml4 alk, fusion, genetics, alk fus...	0.344336
topic1	G	egfr, egfr mutation, gefitinib, epidermal grow...	0.461671
topic3	E	imatinib, gastrointestinal, gist, stromal, mes...	0.875138



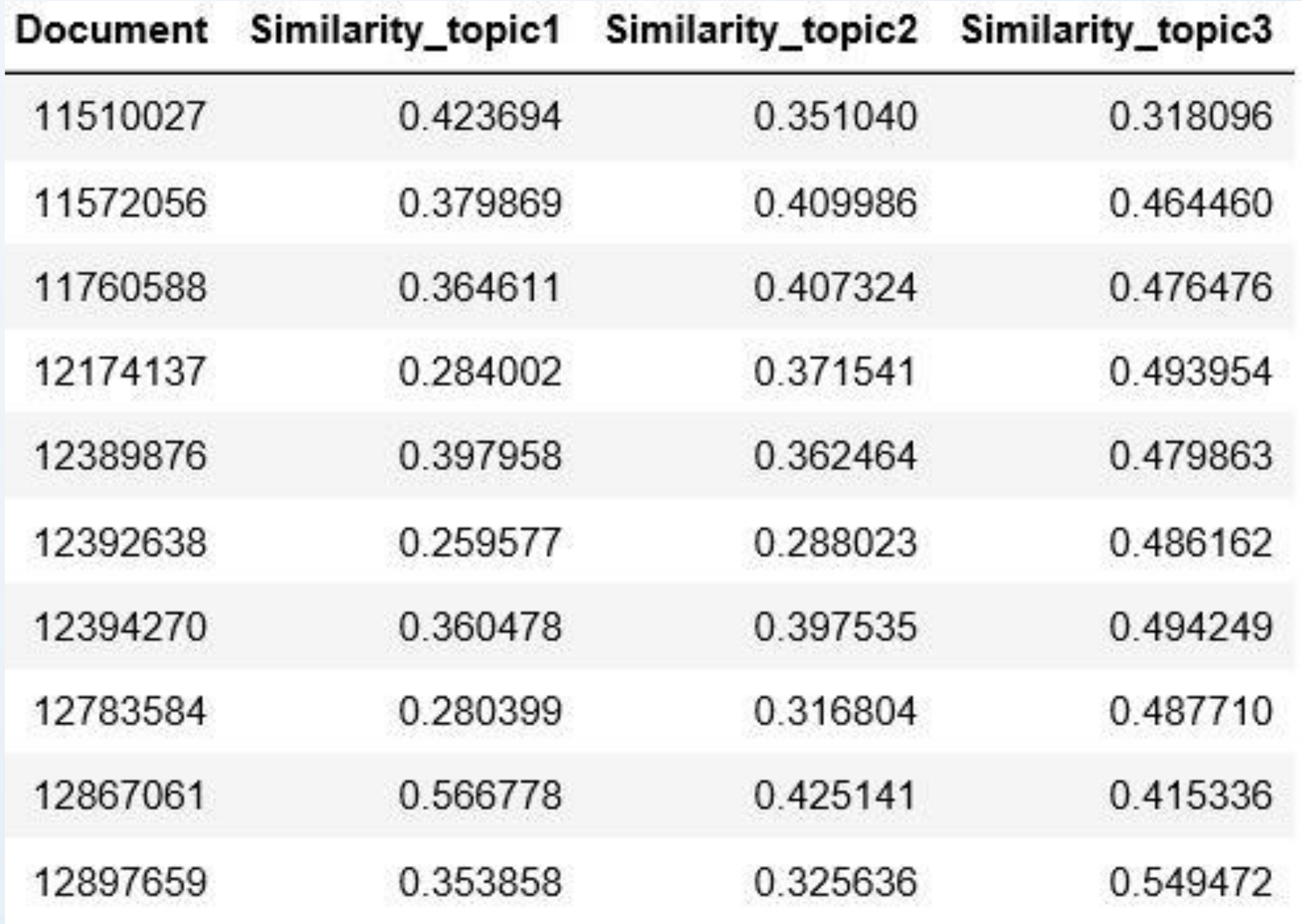


Document	Cluster	Words in Cluster	Distance to Cluster
ASCO_191981-199	G	egfr, egfr mutation, gefitinib, epidermal grow...	0.077840
16503086	G	egfr, egfr mutation, gefitinib, epidermal grow...	0.101019
26798590	F	alk, eml4, eml4 alk, fusion, genetics, alk fus...	0.101703
18509184	G	egfr, egfr mutation, gefitinib, epidermal grow...	0.107333
17196360	E	imatinib, gastrointestinal, gist, stromal, mes...	0.112117
22119437	G	egfr, egfr mutation, gefitinib, epidermal grow...	0.113945
21757253	F	alk, eml4, eml4 alk, fusion, genetics, alk fus...	0.114484
17661208	E	imatinib, gastrointestinal, gist, stromal, mes...	0.115238
AACR_2016-269	G	egfr, egfr mutation, gefitinib, epidermal grow...	0.117347
24916999	G	egfr, egfr mutation, gefitinib, epidermal grow...	0.119828

Documents Ranking on Distance to Cluster

Documents Similarity

Semantic Similarity



Document	Similarity_topic1	Similarity_topic2	Similarity_topic3
11510027	0.423694	0.351040	0.318096
11572056	0.379869	0.409986	0.464460
11760588	0.364611	0.407324	0.476476
12174137	0.284002	0.371541	0.493954
12389876	0.397958	0.362464	0.479863
12392638	0.259577	0.288023	0.486162
12394270	0.360478	0.397535	0.494249
12783584	0.280399	0.316804	0.487710
12867061	0.566778	0.425141	0.415336
12897659	0.353858	0.325636	0.549472

Similarity Score Between a Document and a Query Topic

- Topic 1 (***lung cancer, egfr, aged female***)
- Topic 2 (***lung cancer, eml4-alk, aged male***)
- Topic 3 (***gist, kit exon, aged female***)

Thank You!