Status Report: Research project

Norbert Eke

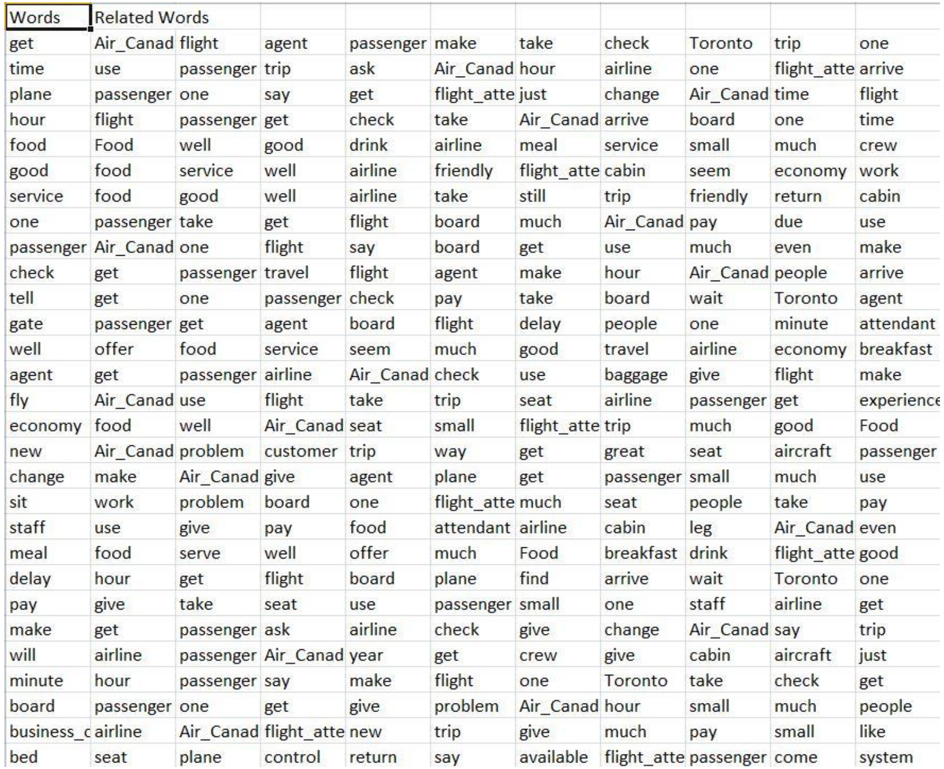
June 21st 2016

Overview

1. Catching up on Abdallah meeting
2. Deep Learning Results
3. Similarity Clustering
4. Topic Word Filtration
5. Likelihood

Deep Learning - (see datasheets)

Each row is a word vector space

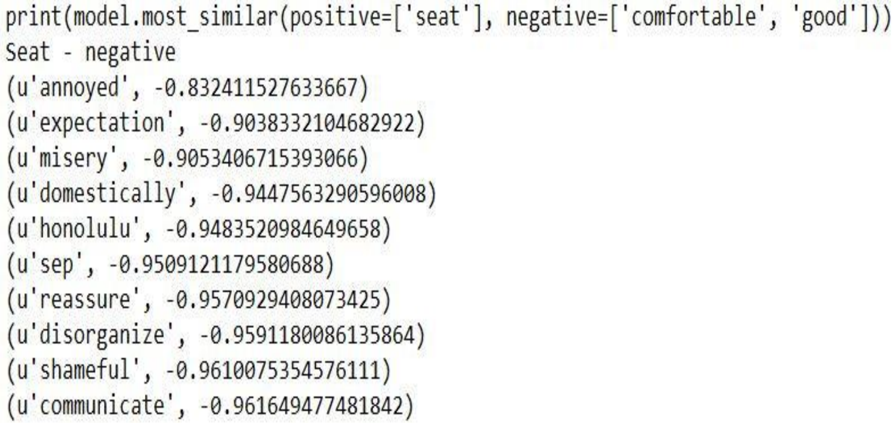


Words	Related Words									
get	Air_Canad	flight	agent	passenger	make	take	check	Toronto	trip	one
time	use	passenger	trip	ask	Air_Canad	hour	airline	one	flight_atte	arrive
plane	passenger	one	say	get	flight_atte	just	change	Air_Canad	time	flight
hour	flight	passenger	get	check	take	Air_Canad	arrive	board	one	time
food	Food	well	good	drink	airline	meal	service	small	much	crew
good	food	service	well	airline	friendly	flight_atte	cabin	seem	economy	work
service	food	good	well	airline	take	still	trip	friendly	return	cabin
one	passenger	take	get	flight	board	much	Air_Canad	pay	due	use
passenger	Air_Canad	one	flight	say	board	get	use	much	even	make
check	get	passenger	travel	flight	agent	make	hour	Air_Canad	people	arrive
tell	get	one	passenger	check	pay	take	board	wait	Toronto	agent
gate	passenger	get	agent	board	flight	delay	people	one	minute	attendant
well	offer	food	service	seem	much	good	travel	airline	economy	breakfast
agent	get	passenger	airline	Air_Canad	check	use	baggage	give	flight	make
fly	Air_Canad	use	flight	take	trip	seat	airline	passenger	get	experience
economy	food	well	Air_Canad	seat	small	flight_atte	trip	much	good	Food
new	Air_Canad	problem	customer	trip	way	get	great	seat	aircraft	passenger
change	make	Air_Canad	give	agent	plane	get	passenger	small	much	use
sit	work	problem	board	one	flight_atte	much	seat	people	take	pay
staff	use	give	pay	food	attendant	airline	cabin	leg	Air_Canad	even
meal	food	serve	well	offer	much	Food	breakfast	drink	flight_atte	good
delay	hour	get	flight	board	plane	find	arrive	wait	Toronto	one
pay	give	take	seat	use	passenger	small	one	staff	airline	get
make	get	passenger	ask	airline	check	give	change	Air_Canad	say	trip
will	airline	passenger	Air_Canad	year	get	crew	give	cabin	aircraft	just
minute	hour	passenger	say	make	flight	one	Toronto	take	check	get
board	passenger	one	get	give	problem	Air_Canad	hour	small	much	people
business_c	airline	Air_Canad	flight_atte	new	trip	give	much	pay	small	like
bed	seat	plane	control	return	say	available	flight_atte	passenger	come	system

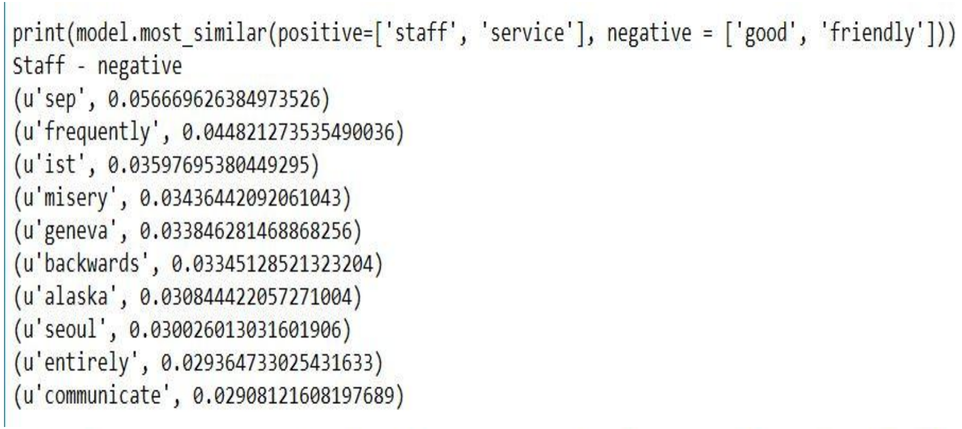
Questions

- Any kind of validation for unsupervised learning ?
- Any way to filter out unrelated things ?
- What can we do with these vector spaces ?
- Doc2Vec vs Word2Vec

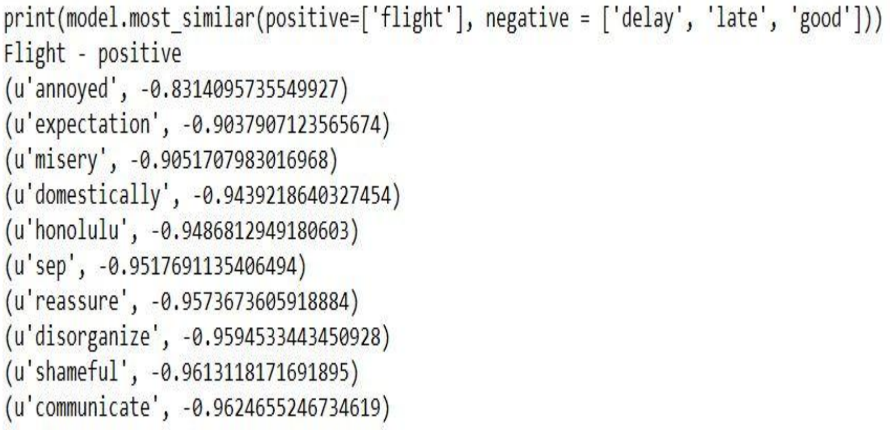
Tried to find what was positive & Negative (beta



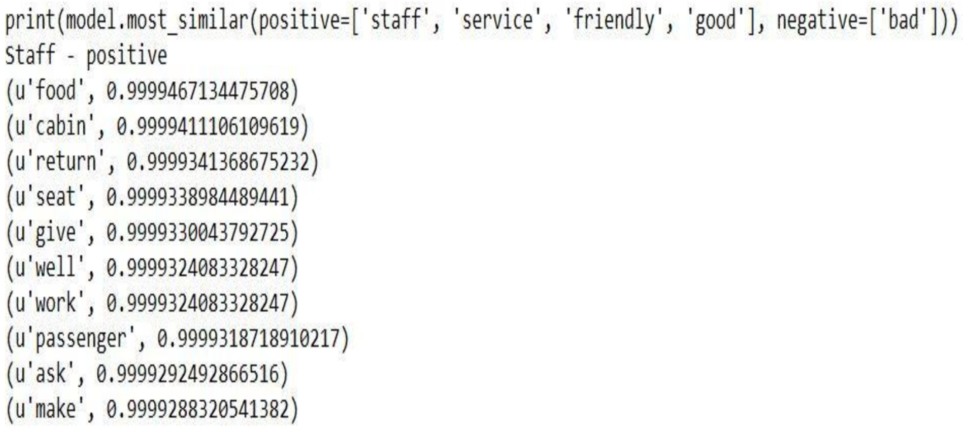
```
print(model.most_similar(positive=['seat'], negative=['comfortable', 'good']))
Seat - negative
(u'annoyed', -0.832411527633667)
(u'expectation', -0.9038332104682922)
(u'misery', -0.9053406715393066)
(u'domestically', -0.9447563290596008)
(u'honolulu', -0.9483520984649658)
(u'sep', -0.9509121179580688)
(u'reassure', -0.9570929408073425)
(u'disorganize', -0.9591180086135864)
(u'shameful', -0.9610075354576111)
(u'communicate', -0.961649477481842)
```



```
print(model.most_similar(positive=['staff', 'service'], negative = ['good', 'friendly']))
Staff - negative
(u'sep', 0.056669626384973526)
(u'frequently', 0.044821273535490036)
(u'ist', 0.03597695380449295)
(u'misery', 0.03436442092061043)
(u'geneva', 0.033846281468868256)
(u'backwards', 0.03345128521323204)
(u'alaska', 0.030844422057271004)
(u'seoul', 0.030026013031601906)
(u'entirely', 0.029364733025431633)
(u'communicate', 0.02908121608197689)
```



```
print(model.most_similar(positive=['flight'], negative = ['delay', 'late', 'good']))
Flight - positive
(u'annoyed', -0.8314095735549927)
(u'expectation', -0.9037907123565674)
(u'misery', -0.9051707983016968)
(u'domestically', -0.9439218640327454)
(u'honolulu', -0.9486812949180603)
(u'sep', -0.9517691135406494)
(u'reassure', -0.9573673605918884)
(u'disorganize', -0.9594533443450928)
(u'shameful', -0.9613118171691895)
(u'communicate', -0.9624655246734619)
```



```
print(model.most_similar(positive=['staff', 'service', 'friendly', 'good'], negative=['bad']))
Staff - positive
(u'food', 0.9999467134475708)
(u'cabin', 0.9999411106109619)
(u'return', 0.9999341368675232)
(u'seat', 0.9999338984489441)
(u'give', 0.9999330043792725)
(u'well', 0.9999324083328247)
(u'work', 0.9999324083328247)
(u'passenger', 0.9999318718910217)
(u'ask', 0.9999292492866516)
(u'make', 0.9999288320541382)
```

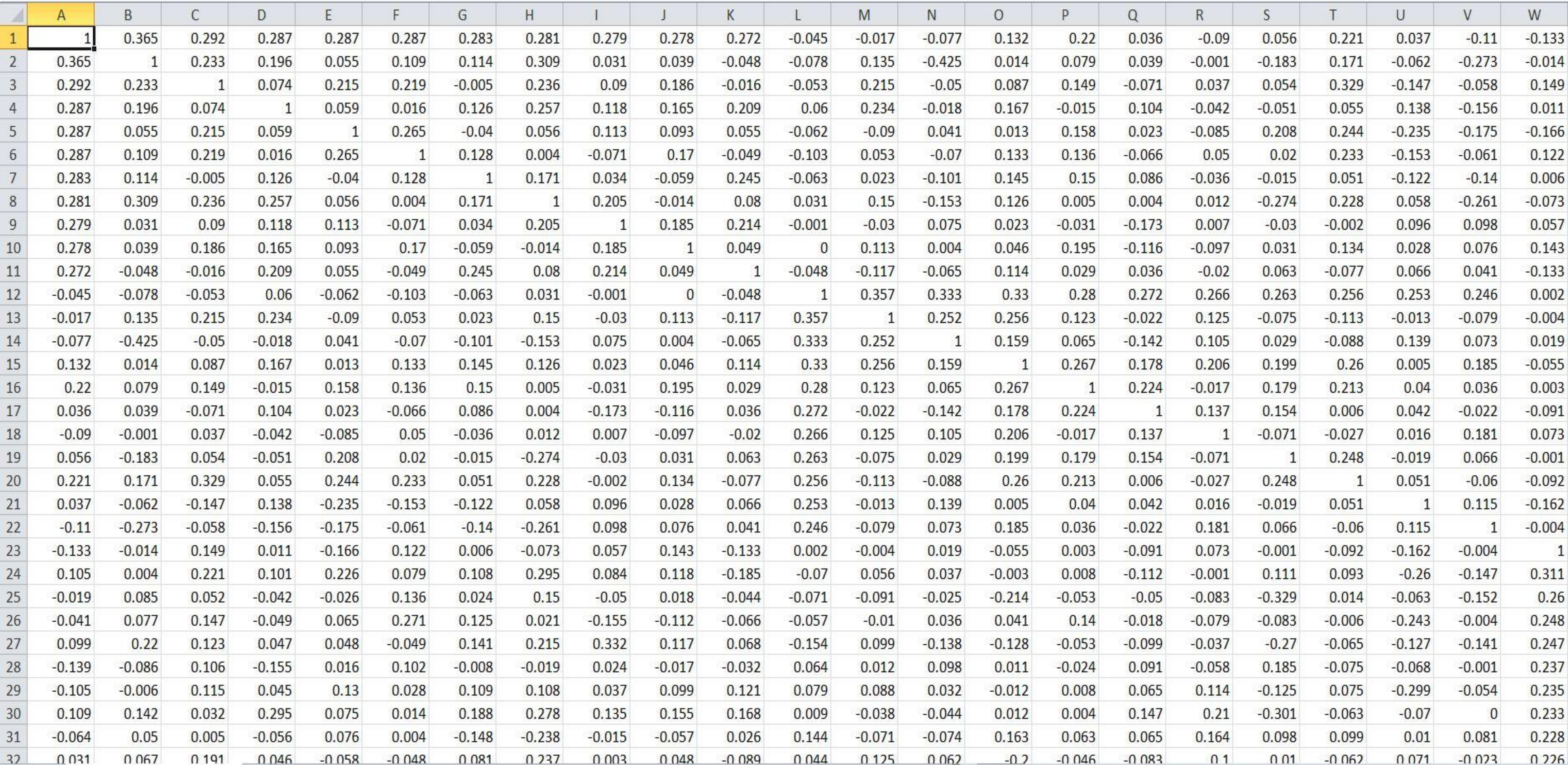
Clustering - see the similarity distance matrices

Ideas

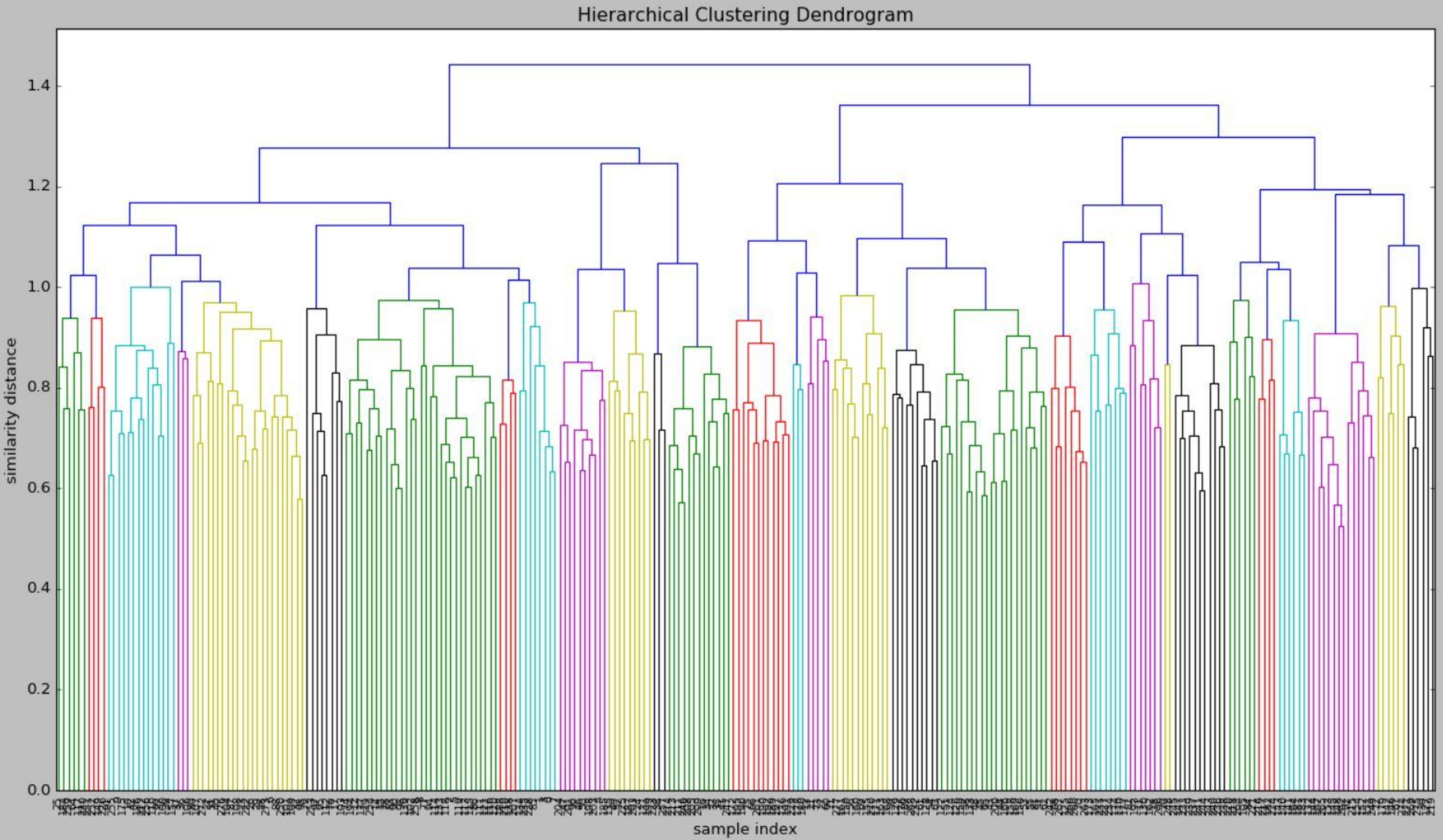
- Word clustering vs word-set(word vector space) clustering
- Would hierarchical work well ?
- Can EM algorithm be applied to the similarity weights

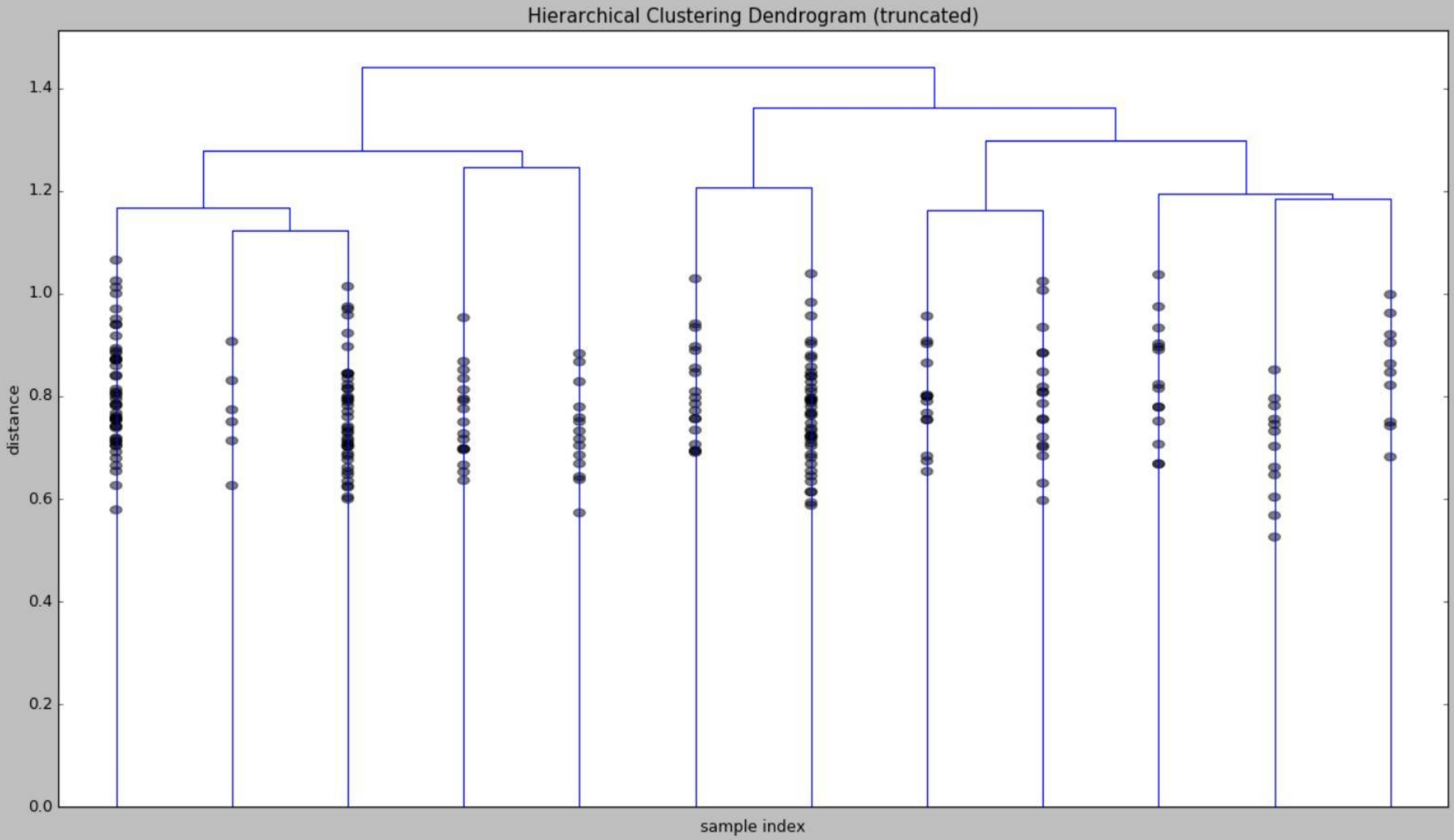
flight attendant
say
plane
take
time
service
Air Canada
food
get
flight
give
trip
good
delay
seat
one
airline
hour
use
passenger

278 x 278 cosine similarity distance matrix on individual words

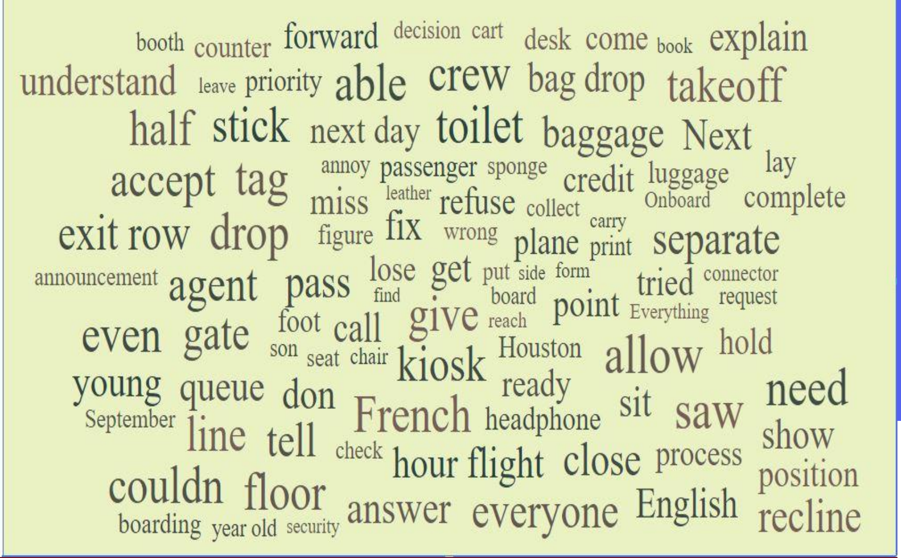


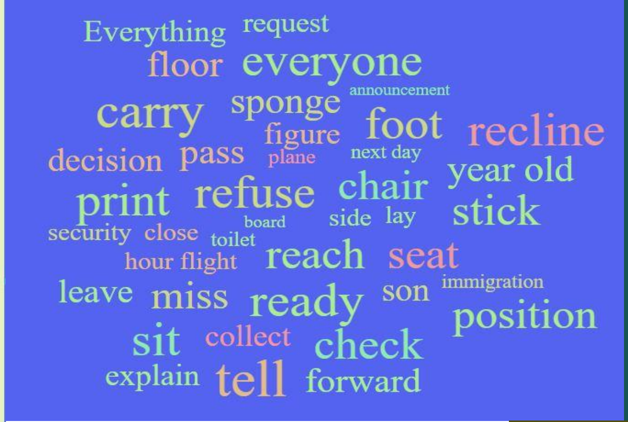
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	1	0.365	0.292	0.287	0.287	0.287	0.283	0.281	0.279	0.278	0.272	-0.045	-0.017	-0.077	0.132	0.22	0.036	-0.09	0.056	0.221	0.037	-0.11	-0.133
2	0.365	1	0.233	0.196	0.055	0.109	0.114	0.309	0.031	0.039	-0.048	-0.078	0.135	-0.425	0.014	0.079	0.039	-0.001	-0.183	0.171	-0.062	-0.273	-0.014
3	0.292	0.233	1	0.074	0.215	0.219	-0.005	0.236	0.09	0.186	-0.016	-0.053	0.215	-0.05	0.087	0.149	-0.071	0.037	0.054	0.329	-0.147	-0.058	0.149
4	0.287	0.196	0.074	1	0.059	0.016	0.126	0.257	0.118	0.165	0.209	0.06	0.234	-0.018	0.167	-0.015	0.104	-0.042	-0.051	0.055	0.138	-0.156	0.011
5	0.287	0.055	0.215	0.059	1	0.265	-0.04	0.056	0.113	0.093	0.055	-0.062	-0.09	0.041	0.013	0.158	0.023	-0.085	0.208	0.244	-0.235	-0.175	-0.166
6	0.287	0.109	0.219	0.016	0.265	1	0.128	0.004	-0.071	0.17	-0.049	-0.103	0.053	-0.07	0.133	0.136	-0.066	0.05	0.02	0.233	-0.153	-0.061	0.122
7	0.283	0.114	-0.005	0.126	-0.04	0.128	1	0.171	0.034	-0.059	0.245	-0.063	0.023	-0.101	0.145	0.15	0.086	-0.036	-0.015	0.051	-0.122	-0.14	0.006
8	0.281	0.309	0.236	0.257	0.056	0.004	0.171	1	0.205	-0.014	0.08	0.031	0.15	-0.153	0.126	0.005	0.004	0.012	-0.274	0.228	0.058	-0.261	-0.073
9	0.279	0.031	0.09	0.118	0.113	-0.071	0.034	0.205	1	0.185	0.214	-0.001	-0.03	0.075	0.023	-0.031	-0.173	0.007	-0.03	-0.002	0.096	0.098	0.057
10	0.278	0.039	0.186	0.165	0.093	0.17	-0.059	-0.014	0.185	1	0.049	0	0.113	0.004	0.046	0.195	-0.116	-0.097	0.031	0.134	0.028	0.076	0.143
11	0.272	-0.048	-0.016	0.209	0.055	-0.049	0.245	0.08	0.214	0.049	1	-0.048	-0.117	-0.065	0.114	0.029	0.036	-0.02	0.063	-0.077	0.066	0.041	-0.133
12	-0.045	-0.078	-0.053	0.06	-0.062	-0.103	-0.063	0.031	-0.001	0	-0.048	1	0.357	0.333	0.33	0.28	0.272	0.266	0.263	0.256	0.253	0.246	0.002
13	-0.017	0.135	0.215	0.234	-0.09	0.053	0.023	0.15	-0.03	0.113	-0.117	0.357	1	0.252	0.256	0.123	-0.022	0.125	-0.075	-0.113	-0.013	-0.079	-0.004
14	-0.077	-0.425	-0.05	-0.018	0.041	-0.07	-0.101	-0.153	0.075	0.004	-0.065	0.333	0.252	1	0.159	0.065	-0.142	0.105	0.029	-0.088	0.139	0.073	0.019
15	0.132	0.014	0.087	0.167	0.013	0.133	0.145	0.126	0.023	0.046	0.114	0.33	0.256	0.159	1	0.267	0.178	0.206	0.199	0.26	0.005	0.185	-0.055
16	0.22	0.079	0.149	-0.015	0.158	0.136	0.15	0.005	-0.031	0.195	0.029	0.28	0.123	0.065	0.267	1	0.224	-0.017	0.179	0.213	0.04	0.036	0.003
17	0.036	0.039	-0.071	0.104	0.023	-0.066	0.086	0.004	-0.173	-0.116	0.036	0.272	-0.022	-0.142	0.178	0.224	1	0.137	0.154	0.006	0.042	-0.022	-0.091
18	-0.09	-0.001	0.037	-0.042	-0.085	0.05	-0.036	0.012	0.007	-0.097	-0.02	0.266	0.125	0.105	0.206	-0.017	0.137	1	-0.071	-0.027	0.016	0.181	0.073
19	0.056	-0.183	0.054	-0.051	0.208	0.02	-0.015	-0.274	-0.03	0.031	0.063	0.263	-0.075	0.029	0.199	0.179	0.154	-0.071	1	0.248	-0.019	0.066	-0.001
20	0.221	0.171	0.329	0.055	0.244	0.233	0.051	0.228	-0.002	0.134	-0.077	0.256	-0.113	-0.088	0.26	0.213	0.006	-0.027	0.248	1	0.051	-0.06	-0.092
21	0.037	-0.062	-0.147	0.138	-0.235	-0.153	-0.122	0.058	0.096	0.028	0.066	0.253	-0.013	0.139	0.005	0.04	0.042	0.016	-0.019	0.051	1	0.115	-0.162
22	-0.11	-0.273	-0.058	-0.156	-0.175	-0.061	-0.14	-0.261	0.098	0.076	0.041	0.246	-0.079	0.073	0.185	0.036	-0.022	0.181	0.066	-0.06	0.115	1	-0.004
23	-0.133	-0.014	0.149	0.011	-0.166	0.122	0.006	-0.073	0.057	0.143	-0.133	0.002	-0.004	0.019	-0.055	0.003	-0.091	0.073	-0.001	-0.092	-0.162	-0.004	1
24	0.105	0.004	0.221	0.101	0.226	0.079	0.108	0.295	0.084	0.118	-0.185	-0.07	0.056	0.037	-0.003	0.008	-0.112	-0.001	0.111	0.093	-0.26	-0.147	0.311
25	-0.019	0.085	0.052	-0.042	-0.026	0.136	0.024	0.15	-0.05	0.018	-0.044	-0.071	-0.091	-0.025	-0.214	-0.053	-0.05	-0.083	-0.329	0.014	-0.063	-0.152	0.26
26	-0.041	0.077	0.147	-0.049	0.065	0.271	0.125	0.021	-0.155	-0.112	-0.066	-0.057	-0.01	0.036	0.041	0.14	-0.018	-0.079	-0.083	-0.006	-0.243	-0.004	0.248
27	0.099	0.22	0.123	0.047	0.048	-0.049	0.141	0.215	0.332	0.117	0.068	-0.154	0.099	-0.138	-0.128	-0.053	-0.099	-0.037	-0.27	-0.065	-0.127	-0.141	0.247
28	-0.139	-0.086	0.106	-0.155	0.016	0.102	-0.008	-0.019	0.024	-0.017	-0.032	0.064	0.012	0.098	0.011	-0.024	0.091	-0.058	0.185	-0.075	-0.068	-0.001	0.237
29	-0.105	-0.006	0.115	0.045	0.13	0.028	0.109	0.108	0.037	0.099	0.121	0.079	0.088	0.032	-0.012	0.008	0.065	0.114	-0.125	0.075	-0.299	-0.054	0.235
30	0.109	0.142	0.032	0.295	0.075	0.014	0.188	0.278	0.135	0.155	0.168	0.009	-0.038	-0.044	0.012	0.004	0.147	0.21	-0.301	-0.063	-0.07	0	0.233
31	-0.064	0.05	0.005	-0.056	0.076	0.004	-0.148	-0.238	-0.015	-0.057	0.026	0.144	-0.071	-0.074	0.163	0.063	0.065	0.164	0.098	0.099	0.01	0.081	0.228
32	0.031	0.067	0.191	0.046	-0.058	-0.048	0.081	0.237	0.003	0.048	-0.089	0.044	0.125	0.062	-0.2	-0.046	-0.083	0.1	0.01	-0.062	0.071	-0.023	0.226


Hierarchical Clustering Dendrogram
similarity distance
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
sample index


Hierarchical Clustering Dendrogram (truncated)
distance
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
sample index

Word Clusters

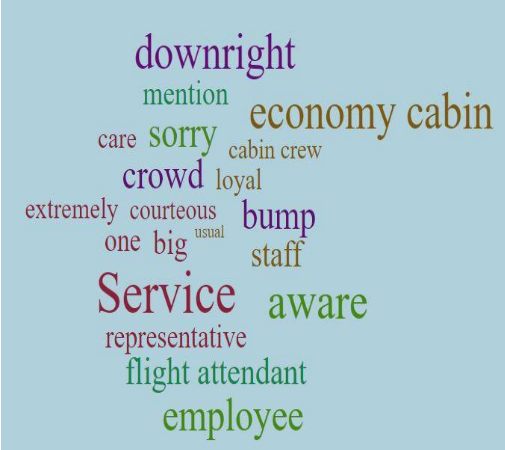


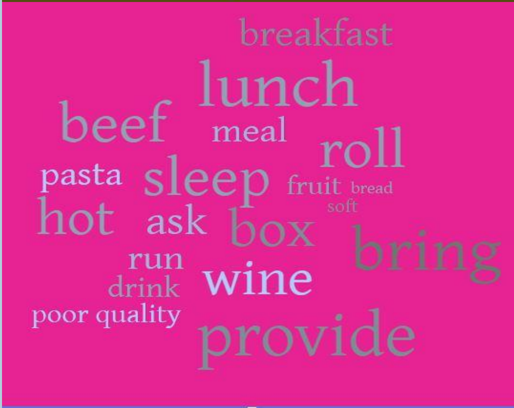


boarding
tried
bag drop
form
Houston
desk
annoy
drop

Despite
cream
technical
delay
eat
nearly
min
minute
land
hour
late
arrive
depart
arrival
promise

luggage
counter
half
queue
understand
headphone
cart
connector
even
allow
call
credit
crew
kiosk
baggage
gate
find
get
French
hold
couldn
put
saw
able
priority
tag
give
lose
line
need
English
wrong
fix
don
agent
passenger
come
separate
accept
answer
process
takeoff
booth



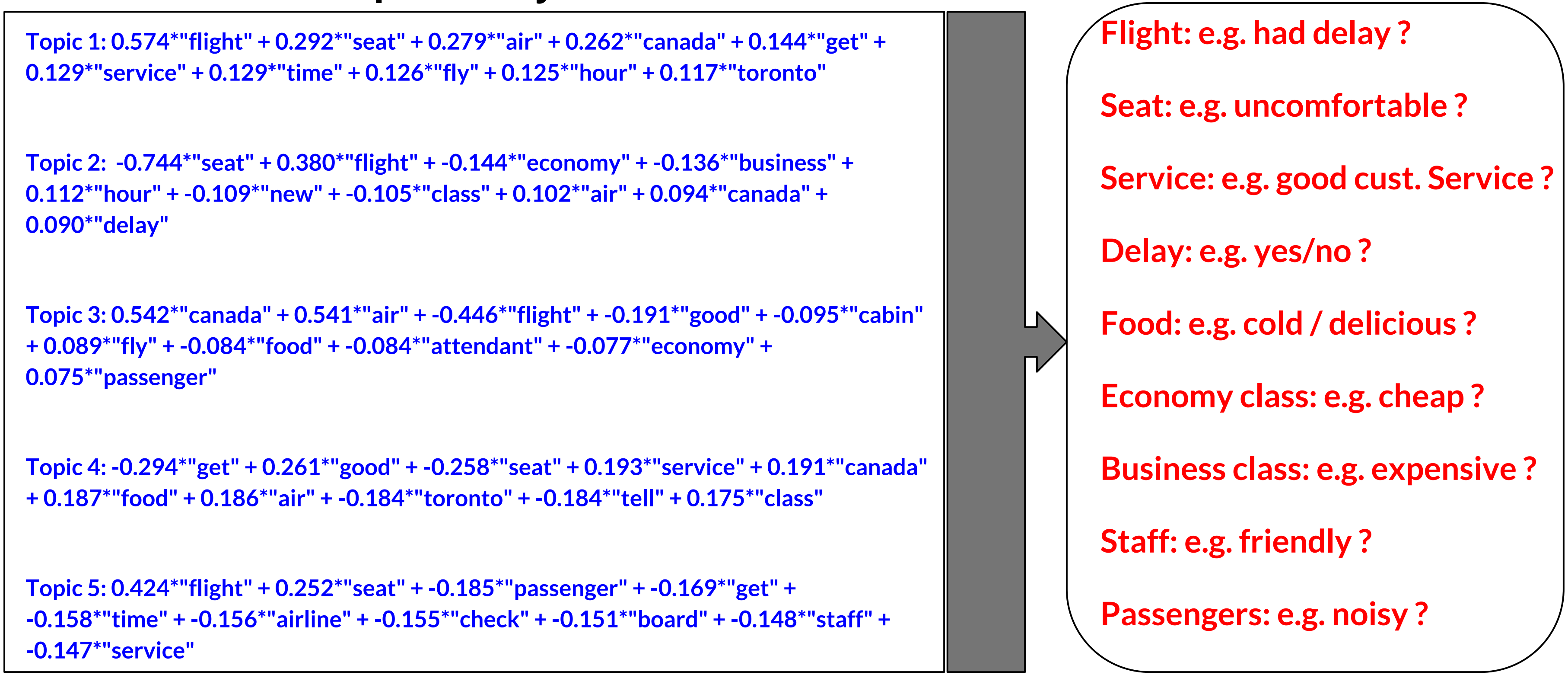


From topic analysis to information retrieval

GOAL: Topic → Information

Topic Analysis

Information retrieved described



Topic 1: 0.574*"flight" + 0.292*"seat" + 0.279*"air" + 0.262*"canada" + 0.144*"get" + 0.129*"service" + 0.129*"time" + 0.126*"fly" + 0.125*"hour" + 0.117*"toronto"

Topic 2: -0.744*"seat" + 0.380*"flight" + -0.144*"economy" + -0.136*"business" + 0.112*"hour" + -0.109*"new" + -0.105*"class" + 0.102*"air" + 0.094*"canada" + 0.090*"delay"

Topic 3: 0.542*"canada" + 0.541*"air" + -0.446*"flight" + -0.191*"good" + -0.095*"cabin" + 0.089*"fly" + -0.084*"food" + -0.084*"attendant" + -0.077*"economy" + 0.075*"passenger"

Topic 4: -0.294*"get" + 0.261*"good" + -0.258*"seat" + 0.193*"service" + 0.191*"canada" + 0.187*"food" + 0.186*"air" + -0.184*"toronto" + -0.184*"tell" + 0.175*"class"

Topic 5: 0.424*"flight" + 0.252*"seat" + -0.185*"passenger" + -0.169*"get" + -0.158*"time" + -0.156*"airline" + -0.155*"check" + -0.151*"board" + -0.148*"staff" + -0.147*"service"

Flight: e.g. had delay ?

Seat: e.g. uncomfortable ?

Service: e.g. good cust. Service ?

Delay: e.g. yes/no ?

Food: e.g. cold / delicious ?

Economy class: e.g. cheap ?

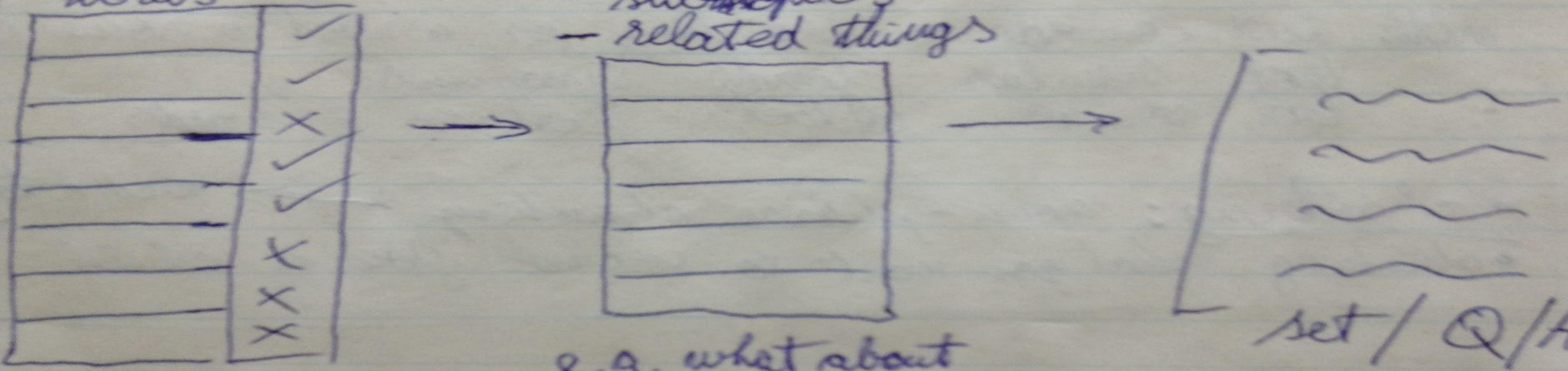
Business class: e.g. expensive ?

Staff: e.g. friendly ?

Passengers: e.g. noisy ?

topic words

- quality/properties OR
- subtopics
- related things



e.g. seat

e.g. what about the seat?

set / Q/A

e.g. description of quality

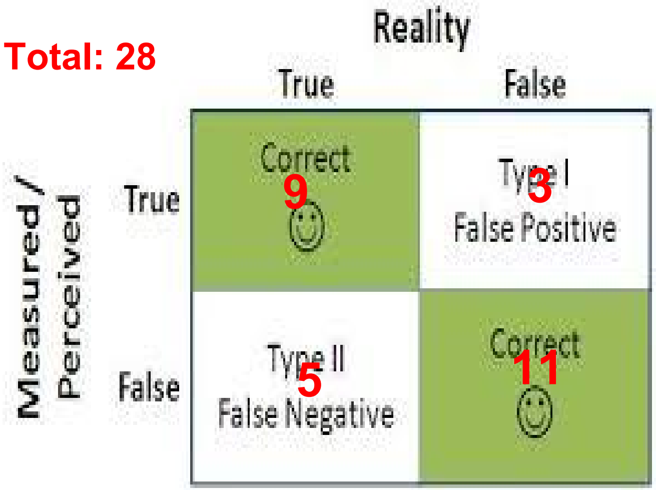
how was it.

- adjective analysis
- some lemmatisation
- frequency TF-iDF
- [illegible] approach

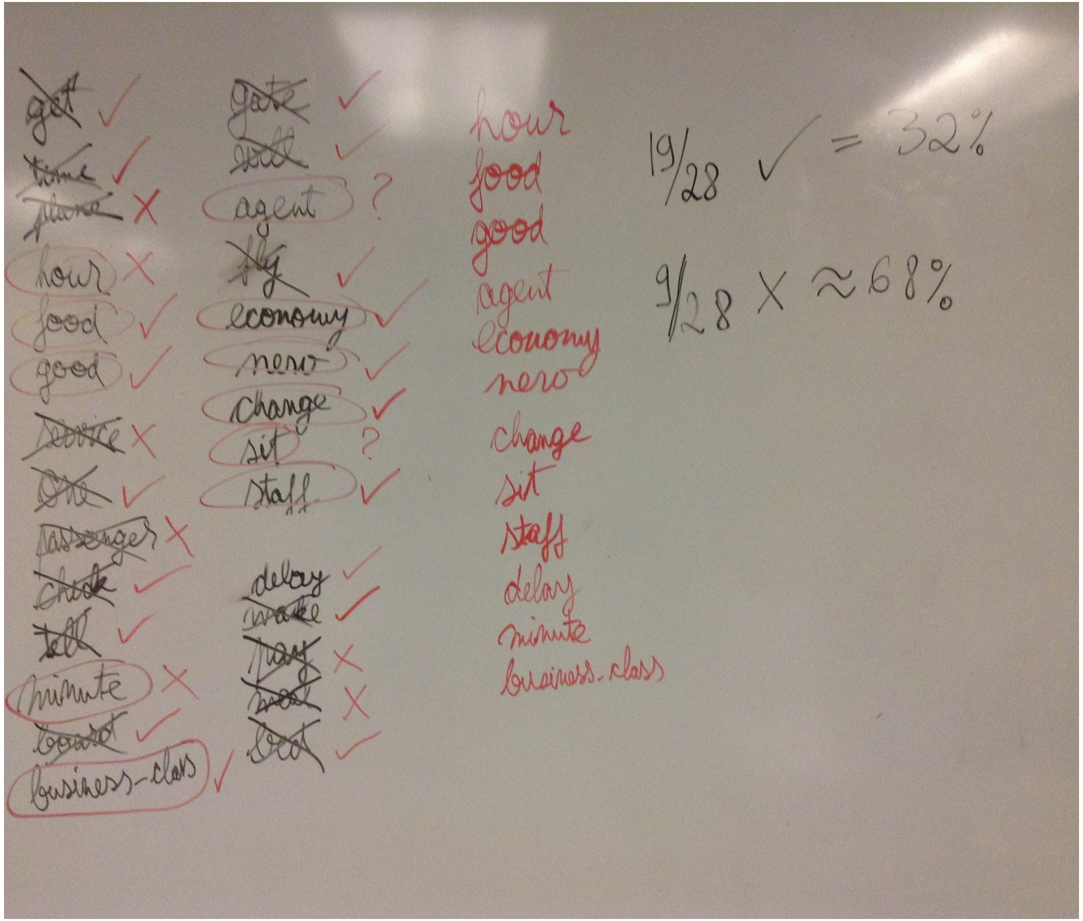
Topic Word Filtering

Throw out unrelated words (garbage)

Total: 28

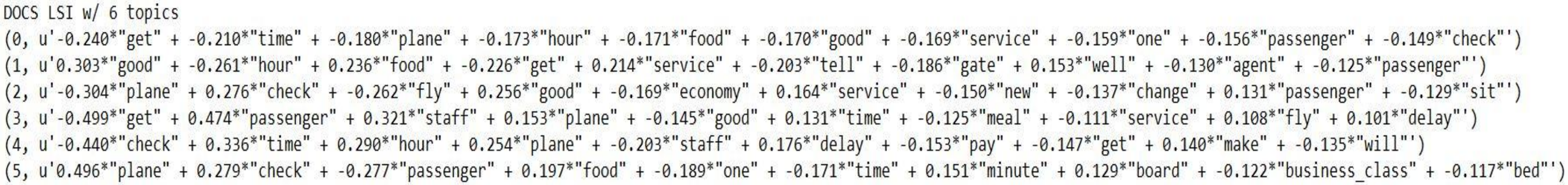


		Reality	
		True	False
Measured / Perceived	True	Correct 9 ☺	Type I 3 False Positive
	False	Type II 5 False Negative	Correct 11 ☺



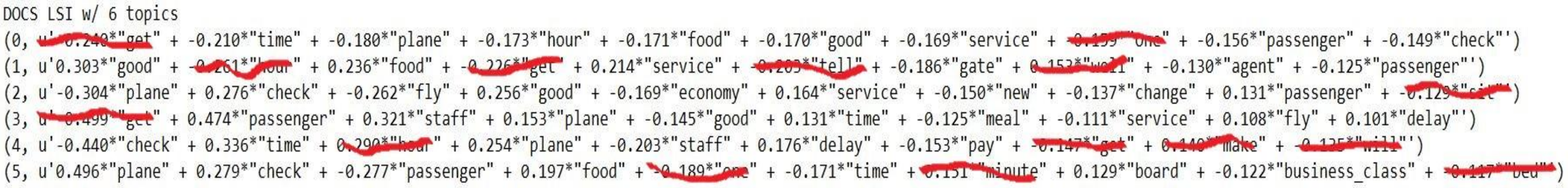
Topic Formatting

A.)



```
DOCS LSI w/ 6 topics
(0, u'-0.240*"get" + -0.210*"time" + -0.180*"plane" + -0.173*"hour" + -0.171*"food" + -0.170*"good" + -0.169*"service" + -0.159*"one" + -0.156*"passenger" + -0.149*"check"')
(1, u'0.303*"good" + -0.261*"hour" + 0.236*"food" + -0.226*"get" + 0.214*"service" + -0.203*"tell" + -0.186*"gate" + 0.153*"well" + -0.130*"agent" + -0.125*"passenger"')
(2, u'-0.304*"plane" + 0.276*"check" + -0.262*"fly" + 0.256*"good" + -0.169*"economy" + 0.164*"service" + -0.150*"new" + -0.137*"change" + 0.131*"passenger" + -0.129*"sit"')
(3, u'-0.499*"get" + 0.474*"passenger" + 0.321*"staff" + 0.153*"plane" + -0.145*"good" + 0.131*"time" + -0.125*"meal" + -0.111*"service" + 0.108*"fly" + 0.101*"delay"')
(4, u'-0.440*"check" + 0.336*"time" + 0.290*"hour" + 0.254*"plane" + -0.203*"staff" + 0.176*"delay" + -0.153*"pay" + -0.147*"get" + 0.140*"make" + -0.135*"will"')
(5, u'0.496*"plane" + 0.279*"check" + -0.277*"passenger" + 0.197*"food" + -0.189*"one" + -0.171*"time" + 0.151*"minute" + 0.129*"board" + -0.122*"business_class" + -0.117*"bed"')
```

B.)



DOCS LSI w/ 6 topics

(0, ~~u'-0.240*"get~~" + -0.210*"time" + -0.180*"plane" + -0.173*"hour" + -0.171*"food" + -0.170*"good" + -0.169*"service" + ~~-0.159*"one~~" + -0.156*"passenger" + -0.149*"check"')

(1, u'0.303*"good" + ~~-0.261*"hour~~" + 0.236*"food" + ~~-0.226*"get~~" + 0.214*"service" + ~~-0.203*"tell"~~ + -0.186*"gate" + ~~0.153*"well~~" + -0.130*"agent" + -0.125*"passenger"')

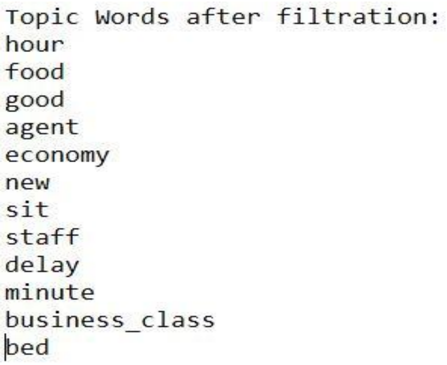
(2, u'-0.304*"plane" + 0.276*"check" + -0.262*"fly" + 0.256*"good" + -0.169*"economy" + 0.164*"service" + -0.150*"new" + -0.137*"change" + 0.131*"passenger" + ~~-0.129*"sit"'~~)

(3, ~~u'-0.499*"get~~" + 0.474*"passenger" + 0.321*"staff" + 0.153*"plane" + -0.145*"good" + 0.131*"time" + -0.125*"meal" + -0.111*"service" + 0.108*"fly" + 0.101*"delay"')

(4, u'-0.440*"check" + 0.336*"time" + ~~0.290*"hour~~" + 0.254*"plane" + -0.203*"staff" + 0.176*"delay" + -0.153*"pay" + ~~-0.147*"get~~" + ~~0.140*"make~~" + ~~-0.135*"will"'~~)

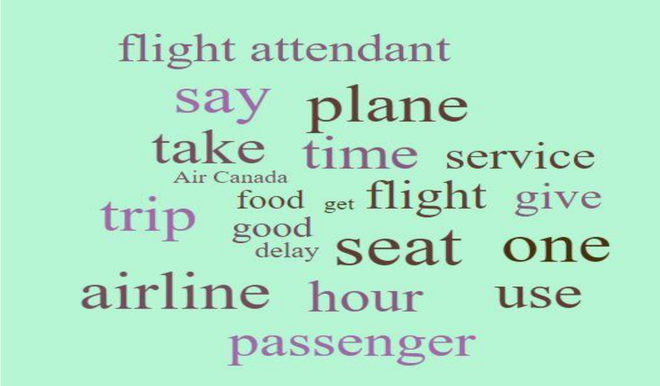
(5, u'0.496*"plane" + 0.279*"check" + -0.277*"passenger" + 0.197*"food" + ~~-0.189*"one~~" + -0.171*"time" + ~~0.151*"minute~~" + 0.129*"board" + -0.122*"business_class" + ~~-0.117*"bed"'~~)

C.)



```
Topic Words after filtration:
hour
food
good
agent
economy
new
sit
staff
delay
minute
business_class
bed
```

D.)



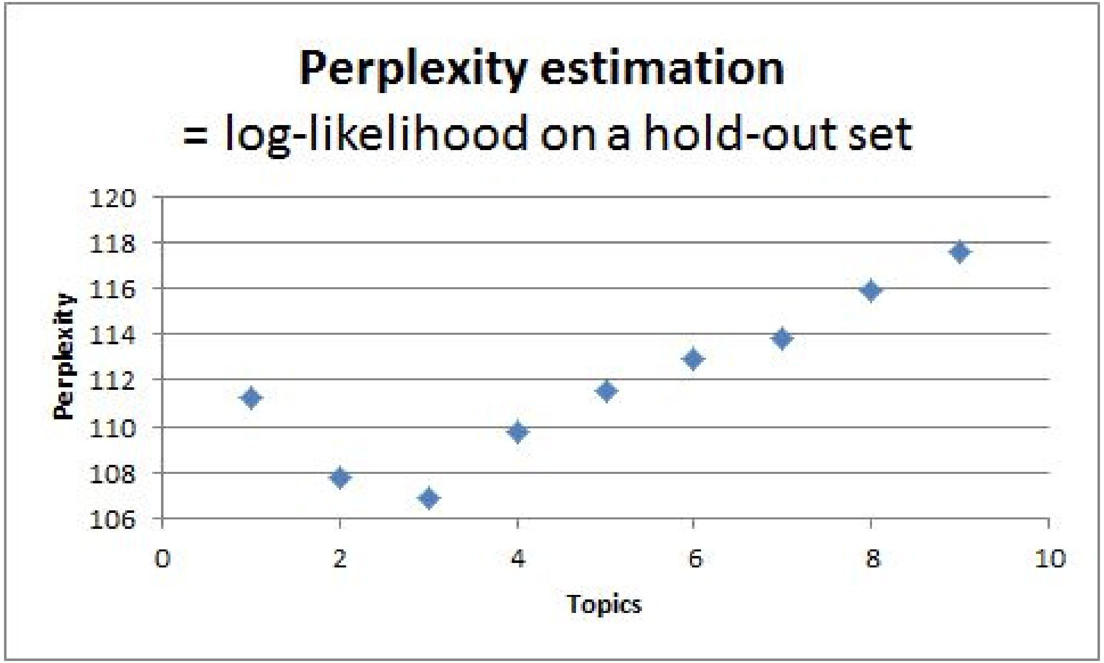
A. Topic terms + coefficients

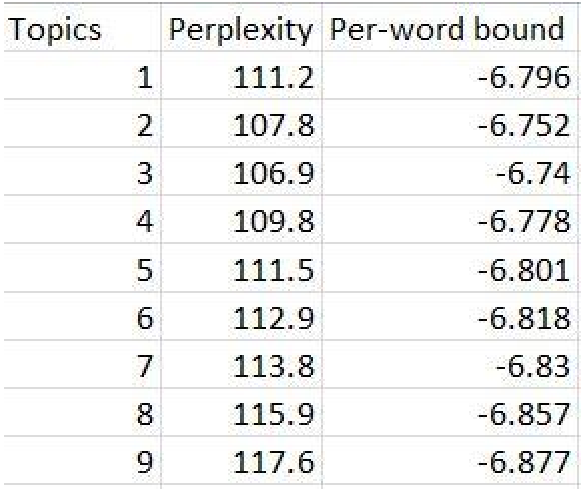
B. A. filtered

C. Filtered keywords, no weights

D. Word Clusters

Likelihood





Topics	Perplexity	Per-word bound
1	111.2	-6.796
2	107.8	-6.752
3	106.9	-6.74
4	109.8	-6.778
5	111.5	-6.801
6	112.9	-6.818
7	113.8	-6.83
8	115.9	-6.857
9	117.6	-6.877

Schedule

June 22 - 25 → Canadian Undergraduate Computer Science Conference

June 27 - July 1 → More Research work

July 5th → Going to Europe

August 18 → Arrive to Kelowna

August 19 - September 1st → Work on Data Interaction Tool