University of Canterbury

Report for building TED recommendation system via Python language

Name: Thong Minh Nguyen (Martin)

Student ID: 68623131

Lecturer: Prof. Christopher Thomson

QUESTION 1:

There are four points of essential criteria in the process of training a topic model for a recommendation engine. At first, to reduce the work and increase the efficiency of training the model, the data from the copora should be cleaned. For example, the new lines and space should be replaced with a standard space, the stop words including popular words in different topic should be removed.

Next, to assess which number of topics is efficient for recommendation engine, the coherence scores script should produce the consistently highest coherence scores. To support the efficiency of this number of topics, the topics listed in the topic distributions of each video/transcript should be checked to make sure they are relevant to each other and to the main content of that video/transcript. Also, the top words in each topic should be relevant and help to name that topic meaningfully, which would prove the semantic stability of the model.

After that, the similarity scores should be used for each recommended videos and sorted out them descendingly. However, the purpose is that to recommend the videos/transcripts which are less relevant, or just averagely similar to the main one, instead of choosing the highest similarity scores. Hence, the lower similarity scores should be considered and the topic distributions of each recommended videos/transcripts should be checked to see if their topics are highly similar to the topics of the main one or not. In particular, to fulfil the criteria, there should be only one topic with the highest/high percentage in recommended videos' topic distributions relating to the given videos. The other topics in these topics distributions should be different or having small quantities with very low percentages if they have the same topics as in the given video's topic distributions.

QUESTION 2:

The preprocessing step is one of the most important stages to increase the efficiency of the recommendation engine.

After developing a function to create a document list that gathers all the transcripts from Ted, the "preprocess_data" function was developed. In this function, at first, the newlines or mutiple sequences of spaces were replaced with a standard space. Next, a loop was created inside the function for each document, to transform all the words into lower case form, tokenise all the texts, remove stop words, and put all final tokens into a list named "texts". Following that, I added some extra stop words ('space', 'love', '000', 'ca') to be removed. These stop words were selected due to the popularity in many topics, and were established from some trials of running the scripts to on the topics' results.

To choose the range of number of topics, another loop was created to calculate the coherence scores for each quantity of topics. In particular, a wide range (0 to 60, interval = 5) was processed at first, and then was reduced down to (20 to 60, interval = 5) as the high coherence scores were focused in this area.

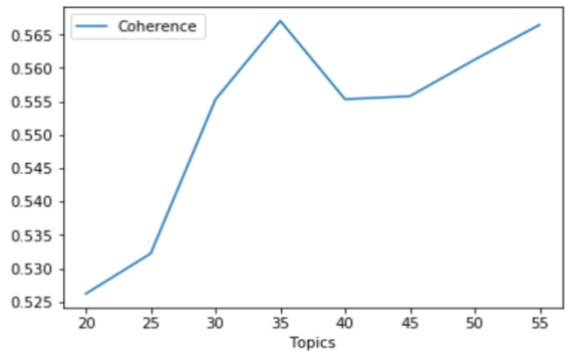

Figure 1. Topic coherence scores graph

After that, the loop was runned through many times to establish the consistent highest score, in this case 35 topics satisfied this condition (as in the Figure 1). For further checking, as to look into the 20 top words list of each topic, they were also clear to be labelled in different categories. After the number of the topics was selected, the training step was processed until each topic could be named meaningfully.

QUESTION 3:

The similarity score and topic distributions would be used at the end of the programming process to find out and evaluate on the recommended videos. As mentioned in the question 1, in lieu of selecting the clips with highest similarity score, the lesser relevant with lower similarity score would be chosen. To explain, the users might be interested in the videos with lower scores that contains other intriguing topics rather than almost the same topic as the video they have just watched. To support this idea, the experiment of Allen, Wheeler-Mackta and Campo on AM/FM radio listeners in Worcester, Massachuset proved that people were likely to select other slightly similar types of music, store in the playlists, and do the selection later rather than keep listening to the same genre (1-2; ch. 1).

Hence, when the similarity score was sorted, the top transcripts having the highest similarity score would be ignore and the next similarity score would be considered, evaluated and tried running serveral times until the suitable recommended transcripts were found.

To begin with, the first video of Kriti Sharma mentioned about the human bias in AI, that she brought up this topics with evidences about current situations and proved that this problem should be avoided to “bring all kind of backgrounds” “to make AI better”. As the resutls from topic distributions, the topics with the highest percentages were technology (number 18 – 18%), gender (number 19 – 14%) and social media (number 22– 11%). This 43% in total was well described the hotly-debated issue Kriti wanted to deliver to the audience – gender bias in digital technology. In other word, this topic model did a good job on categorise the transcripts into well relevant topics that can show the general view of the speaker.

As to the three recommended transcripts of Kriti’s talk, after several trials, I found out that the eighth to the tenth videos would be the good recommended. In particular, they are “how we can teach computers to make sense of our emotions”, “can a robot pass a university entrance exam”, and “Kevin Kelly tells techology’s epic story” with the similarity scores 0.298, 0.293, and 0.287 respectively. Through topic sentences extracted from each transcript, these

documents seemed to have the same topic about technology as the video of Kriti, which satisfied the criteria about being relevant. Additionally, the topic distributions of each recommended documents showed that apart from the main topic (technology – number 18) with highest percentage (15%, 17%, and 21% respectively), the other topics of them were either the same as the least percentage topics in topic distribution of Kriti's video or even not relevant. Hence, this engine worked well on recommending slightly similar to Kriti's transcripts.

The next video of Timothy Bartik was arguing about the postive changes of the American ecomony in the future depending on if and how we could invest in the preschool at the present. As the results from topic distributions of this transcript, the three topics with the highest percentages are economy/business (number 34 – 31%), youth education (number 7 – 20%), and society (number 0 – 16%). To explain, the first and second highest proportion could be well understood that the youth education affects the economy which is relevant to the Timothy's talk. However, the third topic and its top word list seems to be too general and did not match well with the speaker's ideas. Overall, despite unspecific subject, the topic model did relatively good jon on categorising the main opinions of Timothy's talk.

As to the three recommended transcripts of Timothy's video, after some trials, the eighth to tenth documents in descendingly sorted similarity score range were selected. They are "a new kind of job market", "how to escape education's death valley", and "3 stories of local ecoactivism" with the similarity scores as 0.2315, 0.2297, and 0.2202 respectively. As reading through the topic sentences, these recommendations are relevant to the business or education. In addition, as to the topic distributions of each recommended document, apart from one topic with the highest percentage regarding economy/business or education, the other topics are either the same as the least percentage topics in topic distribution of Timothy's talk or not relevant. In other word, the recommendation did a good job as the criteria set before that showed the slightly similar videos.

In conclusion, inspite of some unspecific subjects, the training topic model with 35 topics gave the good results that not only well categorised the main topics of the two given transcripts, but the engine also recommended the slightly similar videos satisfied the criteria set before. However, there were only two given videos used as the testing data for this training topic model, the picture might look different on the big testing dataset. Also to improve the efficiency, another statistical tools should also be applied to increase the semantic interpretation due to the complexity of Ted videos' contents. Hence, another research should be processed to complete this recommendation engine.

References

Allen, David Paul, Henry Jacob Wheeler-Mackta and Jeremy R. Campo. "The Effects of Music Recommendation Engines on the Filter Bubble Phenomenon." *Digital WPI* (2017): 1-15.

DiMaggio, Paul, et al. Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding. Poetics, vol. 41, no. 6, Dec. 2013, pp. 570–606.

Appendix

```
[(0,
  '0.017*"social" + 0.012*"question" + 0.010*"study" + 0.008*"choice" + 0.0
07*"behavior" + 0.007*"group" + 0.007*"human" + 0.006*"reason" + 0.006*"ans
wer" + 0.006*"wrong" + 0.006*"research" + 0.006*"society" + 0.006*"choices"
+ 0.006*"decisions" + 0.006*"questions" + 0.006*"asked" + 0.005*"makes" + 0
.005*"bad" + 0.005*"moral" + 0.005*"evidence"'),
 (1,
  '0.014*"change" + 0.011*"working" + 0.009*"team" + 0.008*"idea" + 0.008*"
community" + 0.007*"create" + 0.007*"ideas" + 0.007*"start" + 0.006*"import
ant" + 0.006*"success" + 0.006*"problems" + 0.006*"process" + 0.006*"organi
zation" + 0.005*"build" + 0.005*"business" + 0.005*"trust" + 0.005*"impact"
+ 0.005*"social" + 0.005*"leaders" + 0.005*"support"'),
 (2,
  '0.026*"black" + 0.016*"white" + 0.014*"law" + 0.013*"police" + 0.011*"ju
stice" + 0.009*"prison" + 0.008*"states" + 0.007*"american" + 0.007*"americ
a" + 0.007*"case" + 0.007*"rights" + 0.007*"crime" + 0.007*"legal" + 0.007*
"court" + 0.006*"community" + 0.006*"race" + 0.005*"violence" + 0.005*"syst
em" + 0.005*"united" + 0.005*"jail"'),
 (3,
  '0.037*"light" + 0.031*"earth" + 0.016*"planet" + 0.015*"universe" + 0.01
2*"sun" + 0.010*"stars" + 0.010*"dark" + 0.010*"mars" + 0.009*"moon" + 0.00
9*"black" + 0.008*"planets" + 0.008*"sky" + 0.007*"solar" + 0.007*"system"
+ 0.007*"star" + 0.007*"big" + 0.007*"hole" + 0.006*"energy" + 0.006*"surfa
ce" + 0.005*"galaxy"'),
 (4,
  '0.040*"children" + 0.035*"family" + 0.026*"mother" + 0.025*"child" + 0.0
18*"home" + 0.017*"parents" + 0.017*"father" + 0.013*"baby" + 0.013*"young"
+ 0.013*"born" + 0.012*"mom" + 0.010*"lives" + 0.009*"families" + 0.009*"gi
rl" + 0.009*"son" + 0.009*"dad" + 0.009*"care" + 0.008*"story" + 0.008*"dau
ghter" + 0.008*"age"'),
 (5,
  '0.049*"percent" + 0.028*"number" + 0.023*"million" + 0.023*"10" + 0.020*
"1" + 0.015*"2" + 0.014*"times" + 0.014*"100" + 0.014*"half" + 0.013*"30" +
0.012*"20" + 0.011*"numbers" + 0.011*"50" + 0.011*"3" + 0.010*"15" + 0.009*
"5" + 0.009*"ago" + 0.008*"average" + 0.008*"40" + 0.007*"answer"'),
 (6,
  '0.017*"political" + 0.017*"government" + 0.017*"power" + 0.013*"country"
+ 0.011*"change" + 0.010*"democracy" + 0.010*"public" + 0.008*"states" + 0.
008*"president" + 0.008*"politics" + 0.007*"united" + 0.006*"citizens" + 0.
006*"national" + 0.006*"state" + 0.006*"america" + 0.005*"vote" + 0.005*"ri
ghts" + 0.005*"society" + 0.005*"media" + 0.005*"election"'),
 (7,
  '0.052*"school" + 0.042*"kids" + 0.029*"students" + 0.023*"education" + 0
.018*"learning" + 0.015*"learn" + 0.014*"high" + 0.013*"children" + 0.012*"
schools" + 0.011*"class" + 0.011*"teachers" + 0.011*"teach" + 0.011*"studen
t" + 0.011*"college" + 0.010*"teacher" + 0.009*"young" + 0.008*"university"
+ 0.008*"teaching" + 0.007*"kid" + 0.006*"skills"'),
 (8,
  '0.031*"cancer" + 0.030*"body" + 0.029*"cells" + 0.019*"heart" + 0.019*"b
lood" + 0.013*"sleep" + 0.012*"cell" + 0.010*"skin" + 0.008*"surgery" + 0.0
07*"tissue" + 0.007*"muscle" + 0.007*"disease" + 0.006*"patient" + 0.006*"s
tem" + 0.005*"bodies" + 0.005*"organs" + 0.005*"tumor" + 0.005*"stress" + 0
.005*"breast" + 0.005*"system"'),
 (9,
  '0.023*"design" + 0.014*"computer" + 0.011*"robot" + 0.011*"build" + 0.00
8*"working" + 0.008*"robots" + 0.008*"system" + 0.008*"create" + 0.007*"mat
erials" + 0.007*"built" + 0.007*"material" + 0.006*"technology" + 0.006*"pa
```

```
per" + 0.006*"designed" + 0.006*"parts" + 0.006*"process" + 0.006*"simple"
+ 0.006*"show" + 0.006*"building" + 0.006*"small"'),
 (10,
  '0.011*"feet" + 0.009*"air" + 0.009*"body" + 0.008*"fly" + 0.008*"move" +
0.008*"foot" + 0.008*"left" + 0.007*"walk" + 0.007*"legs" + 0.006*"ground"
+ 0.006*"side" + 0.006*"end" + 0.006*"mountain" + 0.005*"flying" + 0.005*"p
oint" + 0.005*"running" + 0.005*"walking" + 0.005*"speed" + 0.005*"moving"
+ 0.005*"half"'),
 (11,
  '0.040*"play" + 0.030*"game" + 0.027*"sound" + 0.016*"voice" + 0.015*"vid
eo" + 0.014*"sounds" + 0.014*"games" + 0.013*"playing" + 0.010*"real" + 0.0
10*"hear" + 0.008*"listening" + 0.008*"noise" + 0.008*"hearing" + 0.006*"ex
perience" + 0.006*"call" + 0.006*"players" + 0.006*"start" + 0.006*"win" +
0.005*"player" + 0.005*"played"'),
 (12,
  '0.028*"science" + 0.013*"theory" + 0.009*"universe" + 0.009*"scientists"
+ 0.008*"model" + 0.008*"physics" + 0.007*"question" + 0.007*"scientific" +
0.007*"number" + 0.006*"answer" + 0.006*"quantum" + 0.006*"reality" + 0.006
*"pattern" + 0.006*"understand" + 0.006*"simple" + 0.006*"patterns" + 0.005
*"mathematics" + 0.005*"explain" + 0.005*"matter" + 0.005*"nature"'),
 (13,
  '0.035*"health" + 0.019*"care" + 0.018*"disease" + 0.014*"medical" + 0.01
4*"patients" + 0.011*"drug" + 0.011*"drugs" + 0.010*"doctors" + 0.010*"trea
tment" + 0.009*"doctor" + 0.009*"hospital" + 0.008*"patient" + 0.008*"hiv"
+ 0.007*"medicine" + 0.006*"diseases" + 0.006*"research" + 0.005*"risk" + 0
.005*"early" + 0.005*"test" + 0.005*"healthy"'),
 (14,
  '0.032*"feel" + 0.015*"experience" + 0.012*"feeling" + 0.010*"fear" + 0.0
09*"mind" + 0.008*"person" + 0.008*"lives" + 0.008*"talk" + 0.007*"happy" +
0.007*"live" + 0.007*"happiness" + 0.007*"felt" + 0.007*"friends" + 0.007*"
emotional" + 0.006*"pain" + 0.006*"moment" + 0.006*"emotions" + 0.005*"face
" + 0.005*"experiences" + 0.005*"feelings"'),
 (15,
  '0.024*"show" + 0.019*"video" + 0.014*"film" + 0.012*"story" + 0.011*"pic
ture" + 0.011*"movie" + 0.011*"stories" + 0.011*"real" + 0.010*"audience" +
0.010*"pictures" + 0.009*"camera" + 0.009*"tv" + 0.008*"watch" + 0.007*"man
" + 0.007*"television" + 0.007*"magic" + 0.006*"watching" + 0.006*"face" +
0.006*"started" + 0.006*"movies"'),
 (16,
  '0.054*"food" + 0.019*"eat" + 0.012*"plant" + 0.008*"eating" + 0.008*"fee
d" + 0.007*"plants" + 0.007*"farmers" + 0.006*"bees" + 0.006*"grow" + 0.006
*"sugar" + 0.006*"meat" + 0.006*"farm" + 0.005*"growing" + 0.005*"coffee" +
0.005*"bread" + 0.005*"agriculture" + 0.005*"diet" + 0.005*"fat" + 0.004*"i
nsects" + 0.004*"taste"'),
 (17,
  '0.031*"war" + 0.010*"military" + 0.008*"peace" + 0.008*"conflict" + 0.00
8*"security" + 0.007*"killed" + 0.007*"violence" + 0.006*"country" + 0.006*
"afghanistan" + 0.006*"refugees" + 0.006*"weapons" + 0.005*"soldiers" + 0.0
05*"iraq" + 0.005*"army" + 0.005*"news" + 0.005*"international" + 0.004*"mi
ddle" + 0.004*"east" + 0.004*"refugee" + 0.004*"attack"'),
 (18,
  '0.038*"technology" + 0.024*"human" + 0.021*"car" + 0.018*"future" + 0.01
6*"machine" + 0.011*"cars" + 0.009*"machines" + 0.009*"system" + 0.009*"int
elligence" + 0.009*"problem" + 0.008*"technologies" + 0.007*"power" + 0.007
*"computer" + 0.006*"driving" + 0.006*"humans" + 0.006*"drive" + 0.006*"bui
ld" + 0.005*"ai" + 0.005*"computers" + 0.005*"artificial"'),
 (19,
  '0.084*"women" + 0.040*"men" + 0.026*"woman" + 0.020*"girls" + 0.019*"sex
" + 0.011*"gender" + 0.011*"female" + 0.011*"male" + 0.010*"man" + 0.010*"s
exual" + 0.010*"talk" + 0.009*"young" + 0.008*"girl" + 0.007*"boys" + 0.007
```

```
*"gay" + 0.006*"told" + 0.005*"culture" + 0.005*"talking" + 0.005*"marriage
" + 0.004*"violence"'),
 (20,
  '0.018*"species" + 0.017*"dna" + 0.017*"human" + 0.012*"humans" + 0.011*"
animals" + 0.010*"bacteria" + 0.009*"genes" + 0.008*"evolution" + 0.008*"ge
netic" + 0.008*"animal" + 0.007*"biology" + 0.007*"gene" + 0.007*"cell" + 0
.006*"genome" + 0.006*"living" + 0.005*"ago" + 0.005*"organisms" + 0.005*"f
ound" + 0.005*"molecules" + 0.005*"biological"'),
 (21,
  '0.034*"water" + 0.032*"energy" + 0.017*"climate" + 0.014*"carbon" + 0.01
2*"oil" + 0.012*"air" + 0.011*"change" + 0.009*"power" + 0.008*"waste" + 0.
007*"natural" + 0.007*"gas" + 0.007*"solar" + 0.006*"plastic" + 0.006*"fuel
" + 0.006*"nuclear" + 0.006*"environmental" + 0.006*"planet" + 0.006*"elect
ricity" + 0.006*"trees" + 0.006*"forest"'),
 (22,
  '0.045*"data" + 0.028*"information" + 0.020*"internet" + 0.013*"online" +
0.012*"phone" + 0.010*"technology" + 0.010*"google" + 0.010*"digital" + 0.0
09*"web" + 0.008*"media" + 0.008*"network" + 0.007*"facebook" + 0.006*"open
" + 0.006*"computer" + 0.006*"access" + 0.006*"content" + 0.006*"mobile" +
0.005*"code" + 0.005*"software" + 0.005*"networks"'),
 (23,
  '0.016*"god" + 0.012*"human" + 0.011*"history" + 0.010*"man" + 0.009*"cen
tury" + 0.009*"story" + 0.008*"religion" + 0.007*"ancient" + 0.006*"religio
us" + 0.006*"death" + 0.006*"culture" + 0.006*"faith" + 0.005*"king" + 0.00
4*"nature" + 0.004*"sense" + 0.004*"modern" + 0.004*"live" + 0.004*"dead" +
0.004*"church" + 0.003*"tradition"'),
 (24,
  '0.028*"started" + 0.021*"wanted" + 0.020*"thought" + 0.015*"told" + 0.01
5*"knew" + 0.014*"asked" + 0.014*"found" + 0.013*"decided" + 0.011*"days" +
0.010*"looked" + 0.010*"felt" + 0.010*"realized" + 0.009*"happened" + 0.009
*"months" + 0.009*"night" + 0.009*"learned" + 0.008*"story" + 0.008*"met" +
0.008*"began" + 0.008*"turned"'),
 (25,
  '0.027*"talk" + 0.021*"bit" + 0.019*"idea" + 0.018*"sort" + 0.017*"stuff"
+ 0.016*"problem" + 0.016*"big" + 0.015*"start" + 0.015*"pretty" + 0.014*"i
nteresting" + 0.014*"thinking" + 0.013*"talking" + 0.012*"thought" + 0.012*
"basically" + 0.011*"important" + 0.010*"point" + 0.010*"happen" + 0.009*"h
appened" + 0.009*"couple" + 0.008*"lots"'),
 (26,
  '0.015*"room" + 0.012*"head" + 0.010*"hand" + 0.009*"guy" + 0.008*"eyes"
+ 0.008*"front" + 0.007*"box" + 0.007*"dog" + 0.007*"morning" + 0.007*"feel
" + 0.006*"open" + 0.006*"hard" + 0.006*"night" + 0.006*"person" + 0.006*"h
ands" + 0.006*"door" + 0.006*"kid" + 0.005*"car" + 0.005*"god" + 0.005*"bad
"'),
 (27,
  '0.026*"art" + 0.012*"color" + 0.011*"image" + 0.010*"beautiful" + 0.009*
"project" + 0.009*"artist" + 0.008*"beauty" + 0.008*"idea" + 0.008*"images"
+ 0.007*"museum" + 0.007*"create" + 0.007*"piece" + 0.007*"artists" + 0.006
*"sort" + 0.006*"form" + 0.006*"painting" + 0.006*"series" + 0.006*"experie
nce" + 0.005*"draw" + 0.005*"paint"'),
 (28,
  '0.074*"brain" + 0.012*"memory" + 0.012*"brains" + 0.009*"neurons" + 0.00
8*"information" + 0.007*"human" + 0.007*"consciousness" + 0.006*"body" + 0.
006*"activity" + 0.006*"mind" + 0.006*"control" + 0.005*"visual" + 0.005*"s
ystem" + 0.005*"signals" + 0.005*"eyes" + 0.005*"face" + 0.005*"understand"
+ 0.005*"behavior" + 0.004*"lab" + 0.004*"attention"'),
 (29,
  '0.030*"book" + 0.030*"language" + 0.026*"words" + 0.025*"read" + 0.022*"
word" + 0.018*"write" + 0.017*"story" + 0.017*"books" + 0.015*"english" + 0
.015*"writing" + 0.014*"stories" + 0.010*"reading" + 0.010*"written" + 0.00
```

```
9*"wrote" + 0.007*"speak" + 0.007*"languages" + 0.007*"paper" + 0.006*"lett
ers" + 0.006*"letter" + 0.005*"library"'),
 (30,
  '0.072*"music" + 0.014*"song" + 0.013*"dance" + 0.013*"yeah" + 0.011*"hea
r" + 0.011*"play" + 0.010*"piece" + 0.009*"audience" + 0.009*"singing" + 0.
009*"sound" + 0.008*"la" + 0.007*"playing" + 0.006*"ends" + 0.006*"sing" +
 0.006*"roll" + 0.006*"musical" + 0.006*"listen" + 0.006*"songs" + 0.006*"pi
ano" + 0.005*"ladies"'),
 (31,
  '0.021*"water" + 0.020*"ocean" + 0.015*"fish" + 0.015*"sea" + 0.012*"ice"
 + 0.010*"animals" + 0.007*"species" + 0.007*"deep" + 0.007*"surface" + 0.00
6*"places" + 0.005*"planet" + 0.005*"oceans" + 0.005*"place" + 0.005*"north
" + 0.005*"river" + 0.005*"earth" + 0.005*"found" + 0.005*"coral" + 0.004*"
high" + 0.004*"coast"'),
 (32,
  '0.042*"city" + 0.026*"building" + 0.021*"cities" + 0.013*"design" + 0.01
2*"community" + 0.012*"built" + 0.011*"house" + 0.011*"place" + 0.009*"buil
dings" + 0.009*"build" + 0.009*"live" + 0.009*"places" + 0.009*"public" + 0
.008*"home" + 0.008*"architecture" + 0.008*"york" + 0.008*"street" + 0.007*
"urban" + 0.006*"local" + 0.006*"spaces"'),
 (33,
  '0.029*"countries" + 0.026*"africa" + 0.018*"country" + 0.017*"china" + 0
.015*"india" + 0.014*"global" + 0.009*"growth" + 0.008*"economic" + 0.008*"
poverty" + 0.008*"african" + 0.008*"states" + 0.008*"chinese" + 0.007*"popu
lation" + 0.007*"poor" + 0.007*"united" + 0.007*"percent" + 0.006*"developm
ent" + 0.006*"billion" + 0.006*"south" + 0.006*"income"'),
 (34,
  '0.034*"money" + 0.026*"dollars" + 0.019*"business" + 0.015*"companies" +
 0.014*"company" + 0.013*"market" + 0.011*"buy" + 0.010*"pay" + 0.009*"indus
try" + 0.008*"cost" + 0.008*"jobs" + 0.008*"economy" + 0.007*"percent" + 0.
007*"product" + 0.006*"financial" + 0.006*"sell" + 0.006*"economic" + 0.006
*"million" + 0.006*"costs" + 0.005*"price"')]
```