Styles and Climate Analysis (Proof-of-concept)

KFI research

Note: This POC is to showcase the type of text analysis that we could do to support our studies. I generally relied on this framework here (https://www.tidytextmining.com/ (https://www.tidytextmining.com/)) for the analysis

Before I dive into the modelling aspect of the analysis, I did some exploratory analysis. These are the names of the variables.

```
## [1] "GroupID"    "PerID"      "UserName"   "Version"    "Assessor"  
## [6] "Strengths"  "Weaknesses" "language"  
```

Here is the distribution of **'Version'** variable (in %). **JR's comments: Any idea what's the meaning of this variable?**

```
## 
##    1    2    3    4    5 
## 14.5  0.0 85.5  0.0  0.0 
```

And the distribution of **'Assessor'** variable (in %). **JR's comments: Any idea what's the meaning of this variable?**

```
## 
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15 
## 27.9 13.3 12.3 10.1  8.2  6.0  4.6  3.5  2.8  2.2  1.6  1.3  1.0  0.8  0.7 
##   16   17   18   19   20   21   22   23   24   25   26   27   28   29   30 
##  0.5  0.4  0.4  0.3  0.3  0.2  0.2  0.1  0.1  0.1  0.1  0.1  0.1  0.1  0.1 
##   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45 
##  0.1  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 
##   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60 
##  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 
##   61   62   63   64   65   66   67   68   69   70   71   72   73   74   75 
##  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 
##   76   77   78   79   80   81   82   83   84   85   86   87   88   89   90 
##  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 
##   91   92   93   94   95   96   97   98   99  100  101  102  103  104  105 
##  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 
##  107  108  110  111  112  116  117  119  120  121  122  124  126  134  138 
##  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 
##  144  146  149  153  159  165  169  171  173  176  178  183  184  187  188 
##  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 
##  189  194  197  199  208  214  223  236  244  245  246  261  264  267  270 
##  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 
##  274  284  289  291  294  297  301  302  306  309  313  317  318  320  326 
##  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0 
##  337  338  339  345  347 
##  0.0  0.0  0.0  0.0  0.0 
```

From observation, the data consists of a mixture of languages. Hence, I assigned a language tag to each data point via R's textcat Natural Language Processing package. Here is a distribution (in %) of the languages over 306,127 data points.

```
## 
##          afrikaans           albanian             basque 
##                0.1                0.0                0.2 
##            bosnian             breton            catalan 
##                0.0                0.1                0.9 
##     croatian-ascii     czech-iso8859_2             danish 
##                0.0                0.5                0.4 
##              dutch            english          esperanto 
##                4.7               41.8                0.1 
##           estonian            finnish             french 
##                0.0                0.3                1.5 
##            frisian             german    greek-iso8859-7 
##                0.1                3.4                0.0 
##    hebrew-iso8859_8          hungarian          icelandic 
##                0.0                0.0                0.0 
##          indonesian              irish            italian 
##                0.2                0.0                4.1 
##              latin            latvian         lithuanian 
##                0.2                0.0                0.0 
##              malay               manx     middle_frisian 
##                0.3                0.0                0.3 
##             nepali          norwegian             polish 
##                1.2                0.2                4.3 
##          portuguese           romanian          rumantsch 
##                4.7                0.6                0.2 
##   russian-iso8859_5 russian-windows1251           sanskrit 
##                0.4                0.0                0.1 
##              scots        scots_gaelic      serbian-ascii 
##                0.5               21.1                0.0 
##        slovak-ascii  slovak-windows1250     slovenian-ascii 
##                0.3                0.5                0.0 
## slovenian-iso8859_2            spanish            swahili 
##                0.5                4.9                0.1 
##            swedish            tagalog            turkish 
##                0.2                0.1                0.6 
##    ukrainian-koi8_r              welsh 
##                0.0                0.1
```

Then I limit my analysis to only the data points tagged with 'English Language'. Future developments could involve using Google Translate API to convert the Non-English text to English. But I'm not inclined to do since there're usually subtle nuances associated with each language.

Here is an example of the 1st 2 rows in the dataset (only Strengths used here. I left out 'Weaknesses' column in the analysis).

```
##           GroupID           PerID        UserName
## 8 GROW00000182823 PERW00002545425 MQF552ZH
## 9 GROW00000182459 PERW00002555825 Q3G2L5J5
##
```

```
                                                                Strengths
## 8 She is a very positive and energizing person  She likes getting to know people better and c
ommunicating with them with the tone each will understand better  She likes innovative thinking 
rather than following the standard  Shes a hard worker she feels better when she produces result
s that are valuable  She likes to work wtih teams she believes in the power of teaming  When wor
king on a task with a team she starts from the big picture shares the details later on and makes
sure that everyone understands whats expected of them  She feels happier working with a team rat
her than doing the individual work  Shes a very good speaker She inspires people by telling stor
ies sharing experiences  She believes in what shes doing at Hay Group and this is the source of 
her ultimate satisfaction and she spreds the feeling to her peers around her  For her nothing is
impossible so she does not give up easily she tries all the alternatives at her best  She provid
es clear guidance to her team members motivates them to realize and unleash the potential in the
mselves  She believes in the power of human so she likes spending time on developing people unde
rstanding their needs and expectations and helping the achieve their goals  Shes happy with maki
ng mistakes and learning from her mistakes  For her justice is very important so she pays attent
ion to being fair to her team  She likes to be a role model for the people around her  She is ca
lm under crisis situations and is very innovative with finding solutions to manage people during
those times and solving together with the help of people around her  She likes to empower people
around her and see them improve themselves both personally and careerwise  She likes working wit
h high performers or people with the potential to be so  She loves the process of helping them d
evelop  She takes all the necessary actions to support her team in this respect even improving h
erself to perform better in developing people around her
## 9
```

```
                                Kelly makes you feel like you can bring anything to her and she
will work through it with you  She is very good at letting you bounce ideas around and then givi
ng you a strong direction to take the work away and continue
```

Then I tokenize the dataset through unigrams i.e 1 word / term

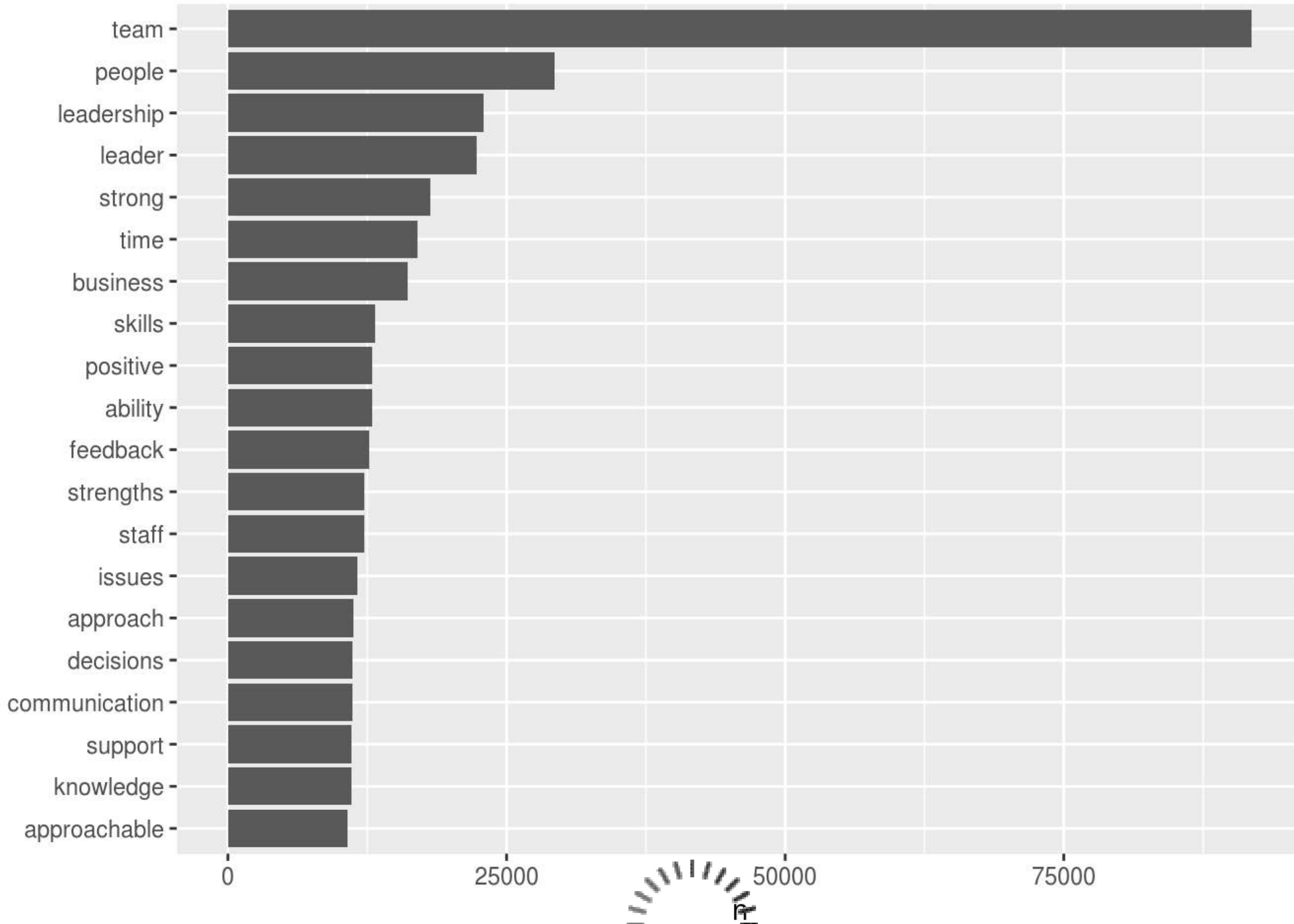
Next I identify stop words in the data frame by merging in a taxonomy from tidytext package. And removed them thereafter.

Count function is used to find the most common words under the 'Strengths Column'.

We could see that the most common words are 'team', 'people', 'leadership', 'positive' - suggesting that people tend to have these attributes as their strenghts

```
## # A tibble: 55,239 x 2
##          word     n
##         <chr> <int>
##  1       team 91888
##  2     people 29322
##  3 leadership 22959
##  4     leader 22328
##  5     strong 18133
##  6       time 17006
##  7   business 16083
##  8     skills 13231
##  9   positive 12961
## 10    ability 12905
## # ... with 55,229 more rows
```

Loading



Loading

Sentiment analysis

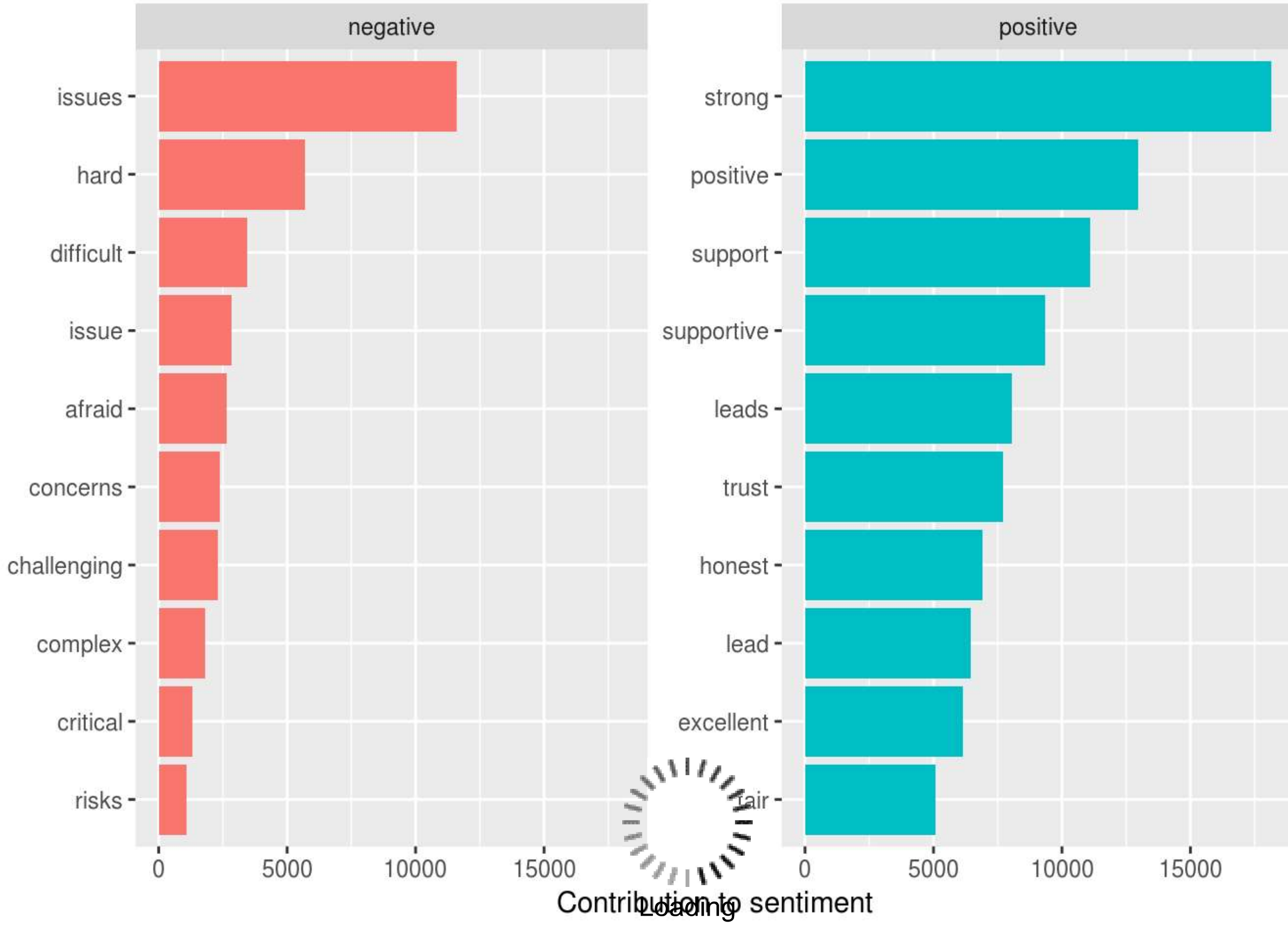
1 of the common toolkits in text analysis is Sentiment Analysis. Using an existing sentiment taxonomy, I'm able to tag words as positive or negative.

We could see that the top 10 most common words are positively connotated which is intuitive since I limited the dataset to only the strengths.

```
## Joining, by = "word"
```

```
## # A tibble: 3,022 x 3
##          word sentiment     n
##         <chr>     <chr> <int>
##  1     strong  positive 18133
##  2   positive  positive 12961
##  3     issues  negative 11608
##  4    support  positive 11115
##  5 supportive  positive  9361
##  6      leads  positive  8066
##  7      trust  positive  7717
##  8     honest  positive  6907
##  9       lead  positive  6445
## 10  excellent  positive  6163
## # ... with 3,012 more rows
```

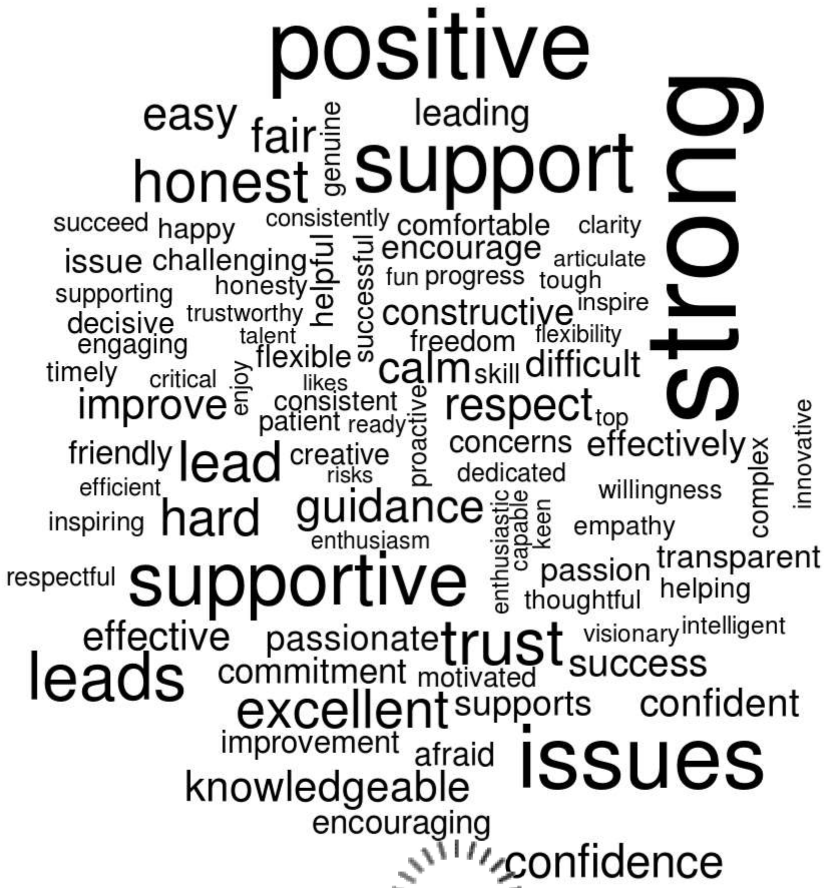
```
## Selecting by n
```



##Word-cloud

A common way to visualize text-analysis is through word clouds. In this section, I plot the word cloud - with the larger words depicting the more common terms in the text.

```
## Loading required package: RColorBrewer
```



TF-IDF (KIV this section first)

Topic-Modelling

Next, I fit a document term matrix into the Latent Dirichlet Allocation (LDA) unsupervised machine-learning framework. ‘Unsupervised Machine Learning’ is just a fancy way of saying Exploratory Thematic Analysis.

As a POC, I set 2 as the number of latent topics that are able to represent the dataset. There are diagnostics to determine the optimal number which I will use in future iterations.

```
## A LDA_VEM topic model with 2 topics.
```

Word-topic probabilities

LDA allows us to extract per-topic-per-word probabilities (Beta) from the model.

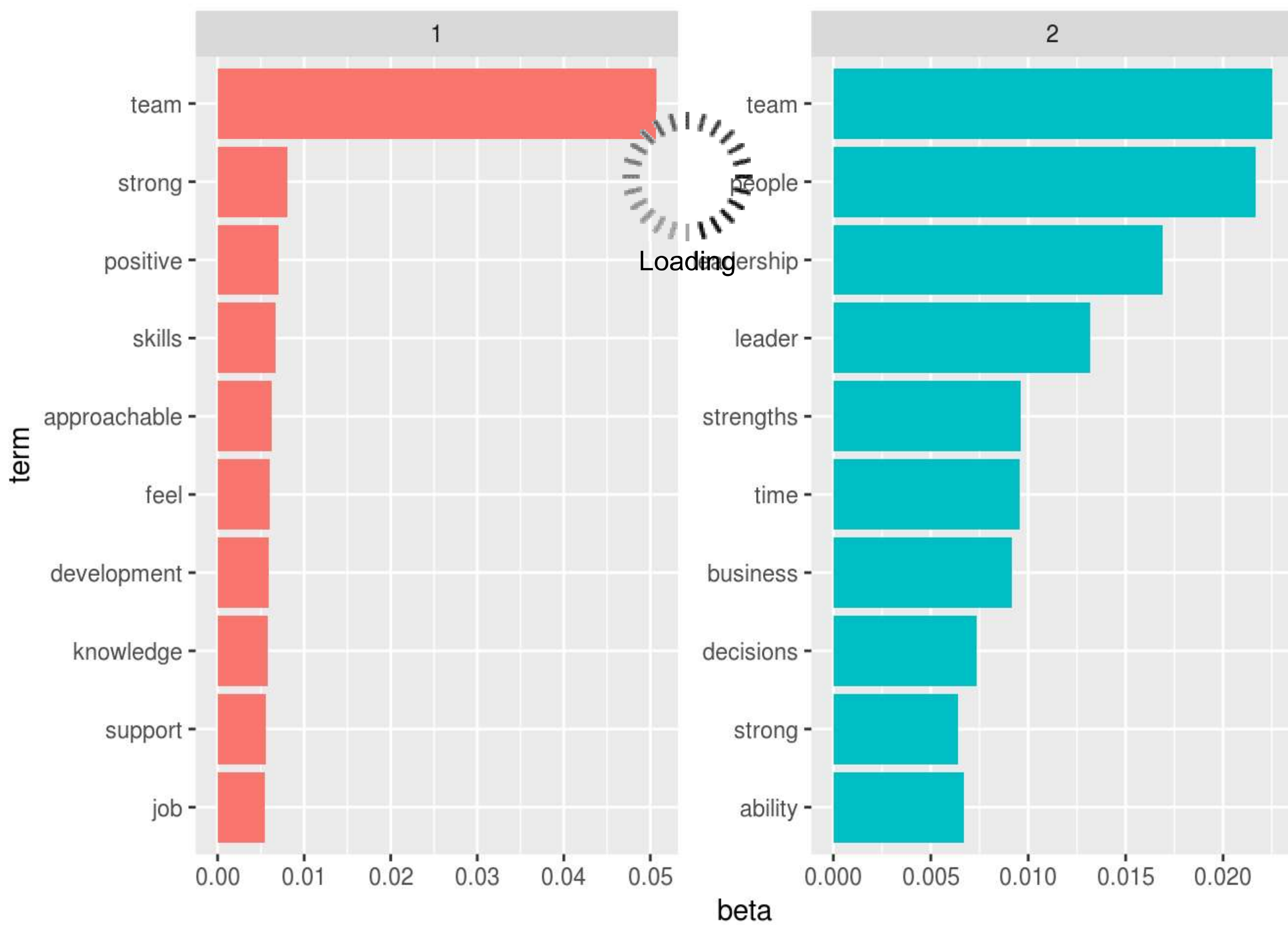
The model assigned probabilities to each term of being generated from either of the 2 topics.

As you notice, many of the terms below are not formatted properly (e.g. remove non-alphabets; keep only the root words through stemming). I will do so in further iterations.

```
## # A tibble: 110,478 x 3
##    topic    term         beta
##    <int>   <chr>        <dbl>
##  1     1      aa 6.368539e-07
##  2     2      aa 1.603523e-07
##  3     1   aaahc 2.951243e-07
##  4     2   aaahc 5.015094e-07
##  5     1   aacqa 5.892588e-07
##  6     2   aacqa 2.078677e-07
##  7     1   aacsb 2.698668e-07
##  8     2   aacsb 5.267246e-07
##  9     1 aadhaar 5.136081e-07
## 10     2 aadhaar 2.833916e-07
## # ... with 110,468 more rows
```

Next, I find the top 10 terms that best represents each topic. In the graphs below, it doesn't show any clear distinction - with a couple of terms falling in both baskets.

Model could be further tuned to obtain better distinctions (e.g. increasing number of topics, further data-cleaning, using n-grams i.e. n words per term instead of 1 word per term now)



Per-document classification

Through LDA, we are also able to assign a topic to each document - in our case, a topic to each comment. In the 1st 2 rows for Documents 2222276Z, we see gammas of 49.3% and 50.7%. This means that an estimated 49.3% comes from Topic 1 and 50.7% from Topic 2.

```
## # A tibble: 250,804 x 3
##          document topic     gamma
##             <chr> <int>     <dbl>
##  1 2222276Z           1 0.4894329
##  2 2222276Z           2 0.5105671
##  3 22224ZTD           1 0.5143616
##  4 22224ZTD           2 0.4856384
##  5 2223455G           1 0.4929482
##  6 2223455G           2 0.5070518
##  7 222365KZ           1 0.5008551
##  8 222365KZ           2 0.4991449
##  9 2223H4KC           1 0.4925733
## 10 2223H4KC           2 0.5074267
## # ... with 250,794 more rows
```

Potential applications

With LDA, we’re able to find out the topics that are commonly associated with high performers. And if there’re demographic and job variables, we can explore these tendencies by gender, occupations, industries, etc.

Loading