Coding problems

1. Actual problem

We have a python API which works with Twilio and Google speech aPI.

At present, the system generates a call recording but is not plugged into Google API.

We need to add functionality so that once a call is complete, the code sends the music file to Google Speech API and then stores it in the TRANSCRIPTION part of the database so that our main website can fetch it (using a CRON job)

2. Possible problem

We want to be able to create a keyword cloud for experts whereby you can enter NAME and COMPANY.

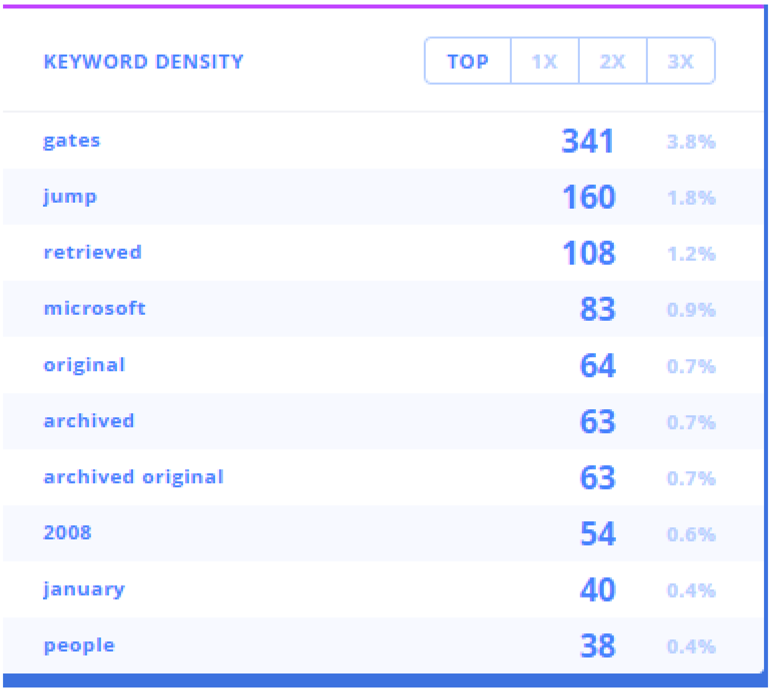
A script should then thread this string ‘Bill Gates Microsoft’ through Google search and then scrape the first X search pages and the first Y news articles.

It should build up a temporary base of the most common words, bigrams and trigrams.

It should then remove those entries which have non specific or irrelevant words (a, and, a, as as well as common marketing jargon like subscribe now, follow, like this page etc etc)

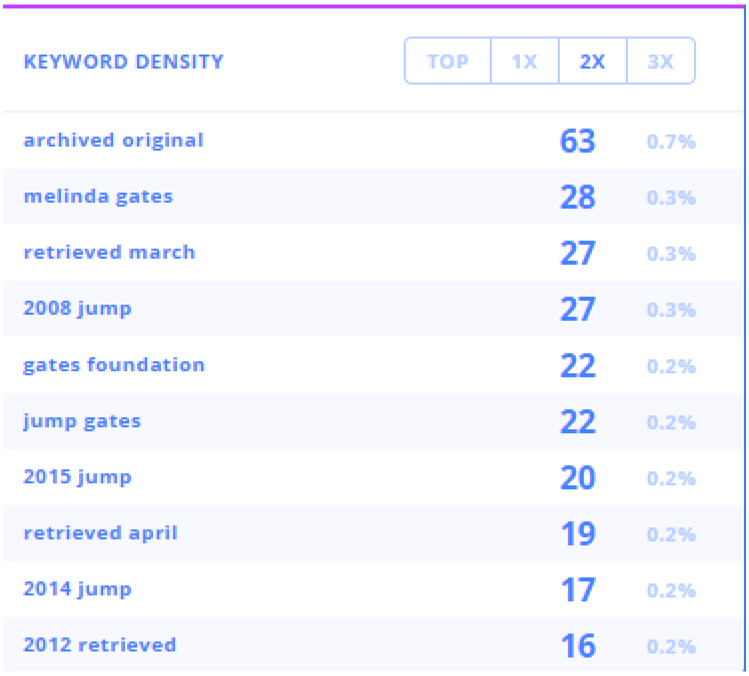
The system should then return a list of single words, bigrams and trigrams that are relevant keywords that defines the person. If we used Bill Gates Microsoft, we should get thing.

I have manually done this by copying the first 5 pages of BILL GATES MICROSOFT into wordcounter.io with the following results



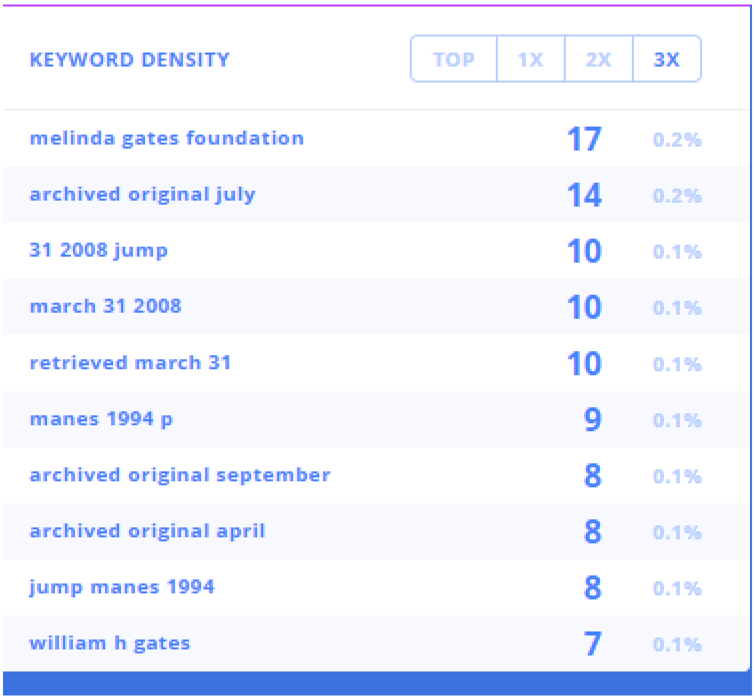
KEYWORD DENSITY

| | | |
|---|---|---|
| gates | 341 | 3.8% |
| jump | 160 | 1.8% |
| retrieved | 108 | 1.2% |
| microsoft | 83 | 0.9% |
| original | 64 | 0.7% |
| archived | 63 | 0.7% |
| archived original | 63 | 0.7% |
| 2008 | 54 | 0.6% |
| january | 40 | 0.4% |
| people | 38 | 0.4% |



KEYWORD DENSITY

| | | |
|---|---|---|
| archived original | 63 | 0.7% |
| melinda gates | 28 | 0.3% |
| retrieved march | 27 | 0.3% |
| 2008 jump | 27 | 0.3% |
| gates foundation | 22 | 0.2% |
| jump gates | 22 | 0.2% |
| 2015 jump | 20 | 0.2% |
| retrieved april | 19 | 0.2% |
| 2014 jump | 17 | 0.2% |
| 2012 retrieved | 16 | 0.2% |



KEYWORD DENSITY

| | | |
|---|---|---|
| melinda gates foundation | 17 | 0.2% |
| archived original july | 14 | 0.2% |
| 31 2008 jump | 10 | 0.1% |
| march 31 2008 | 10 | 0.1% |
| retrieved march 31 | 10 | 0.1% |
| manes 1994 p | 9 | 0.1% |
| archived original september | 8 | 0.1% |
| archived original april | 8 | 0.1% |
| jump manes 1994 | 8 | 0.1% |
| william h gates | 7 | 0.1% |

Obviously you can see there is a lot of irrelevant regular keywords. We need to flush them all out