Knowledge Media Institute

The Open University

Discovery of listening experiences

Phase one (March 2018)

Problem / *Objective*

How to identify accounts of listening experiences from texts?

An automatic bookmarks generator for texts identifying candidates Listening Experiences

Approach

We make the hypothesis that these are a subset of the texts talking about music.

Phase 1:

To develop a dictionary of terms whose occurrence in a text could signify a discourse about music.

To show that this dictionary represents well Listening Experiences (LE) in the database.

Phase 2:

To design/generate a *model* by using the dictionary in combination with *features* of LE.

To develop a system that generate annotations of texts and evaluate it on a gold standard of LE and associated sources.

Today we report on Phase 1.

Development of a music dictionary

- Gutenberg corpus (english subset)
- We NLP to get a vector of terms for each documents
- We calculated TF/IDF of each doc/term pair in Gutenberg
- We collecting the terms in documents classified in the Music shelf
- We sorted them by relevance towards the sub-corpus
- We validated the dictionary against the LED set and the Reuters-21578 corpus (as negative)

Text to vector (NLP)

- Removing stopwords, keeping POS information
- Example: *“So the Rontgens have played you the new Brahms symphony! - another of my few musical joys taken from me! It always happens that when I have been specially counting on something of the sort as regards you, Fate […]”* - LED-1438250799133

```
0  rontgen[NNS]
1  play[VBN]
2  Brahms[NNP]
3  symphony[NN]
4  another[DT]
5  musical[JJ]
6  take[VBN]
7  always[RB]
8  happen[VBZ]
9  specially[RB]
10 count[VBG]
11 something[NN]
12 sort[NN]
13 regard[VBZ]
14 Fate[NNP]
...
```

Knowledge Media Institute

The Open University

TFIDF

"In information retrieval, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect ***how important a word is to a document*** *in a collection or corpus." [1]*

```
term_freq = term_usages / doc_size
idf = LOG(48790 / num_docs_with_term)
tf_idf = term_freq*idf
```

Highest TF-IDF: 1.5901121823585802
Lowest TF-IDF: 4.032538525747152e-08
Highest TF-IDF in the Music Shelf: 0.0922981613222286
Lowest TF-IDF in the Music Shelf: 7.517321708209822e-07

[1] https://en.wikipedia.org/wiki/Tf-idf

```
Document: Gutenberg-15141
—
Beethoven[NNP]      0.072727554032261 93
Symphony[NNP]      0.015139485794100219
Schindler[NNP]      0.007967133189523013
Vienna[NNP]      0.007256378255299395
Haydn[NNP]      0.007141321088599549 5
Wagner[NNP]      0.007088376068171141
Breuning[NNP]      0.006717815731235641
Ries[NNP]      0.006111818988630585
Mozart[NNP]      0.005978596454218494 5
Lichnowsky[NNP]      0.005846276132915727
quartet[NNS]      0.005422433661927 3
Czerny[NNP]      0.005217816538462906
Mass[NNP]      0.005135716029154898
opus[NN]      0.004832913297756952
composer[NN]      0.004442636696952911
Karl[NNP]      0.004326928936343346
Holz[NNP]      0.004142928952284239
Bach[NNP]      0.003742500417903241 7
sonata[NNS]      0.003561838356334826
Bonn[NNP]      0.00355707250098514
symphony[NNS]      0.003447084601144992
music[NN]      0.003265220377074476 8
```

Statistics

- Gutenberg (english): **48790** documents, **79** in the *Music* shelf
- Number of doc/terms occurrences: 1.460.211.421
- Number of distinct terms: 7.183.327
- Number of terms occurring only in 1 doc: 4.405.918
- Number of doc/terms in the Music Shelf: 1.934.581
- Number of distinct terms in the Music Shelf: 89.883
  - **1.25%** of the total of distinct terms in the corpus

Knowledge Media Institute

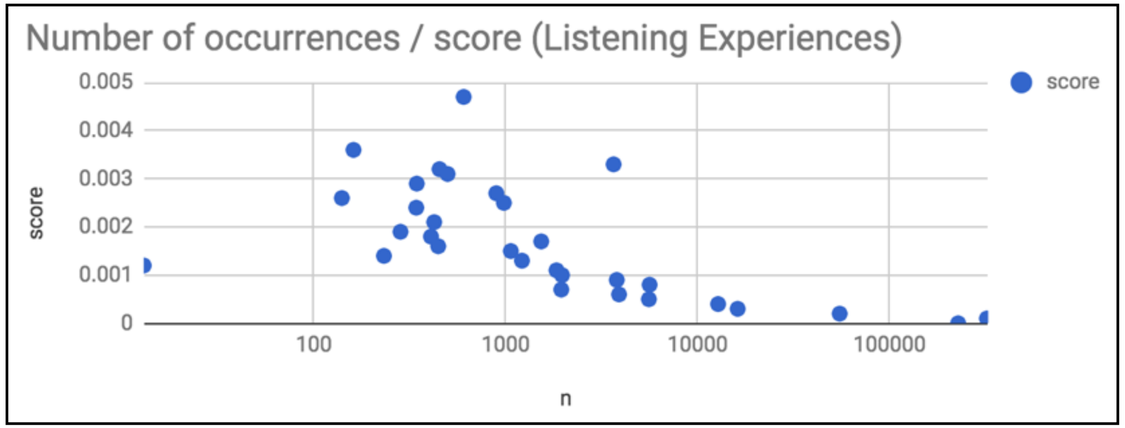
Dictionary

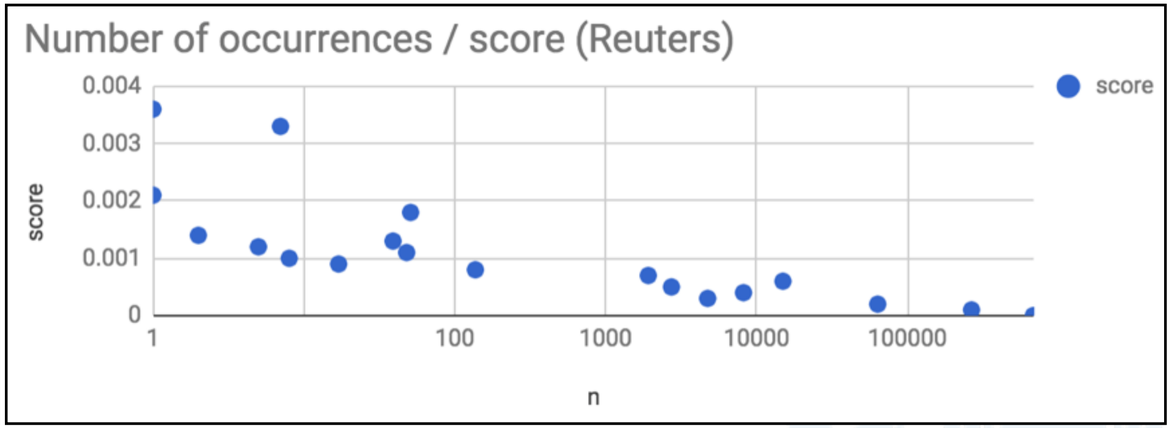
- 89.883 terms ordered by relevance
- Relevance = AVG(TFIDF) of docs in Music Shelf

Beethoven[NNP]	0.004708996602	1
vocal[JJ]	0.003577405412	2
music[NN]	0.003279422105	3
Liszt[NNP]	0.003201453413	4
Chopin[NNP]	0.003163986853	5
composer[NN]	0.003115849809	6
Mozart[NNP]	0.002860199248	7
musical[JJ]	0.002722584954	8
Haydn[NNP]	0.002579207714	9
piano[NN]	0.002500942374	10
aria[NN]	0.0006770586871	98
fugue[NN]	0.0006655704232	99
theme[NN]	0.0006590153165	100
accent[NN]	0.000222760115	497
master[NNS]	0.0002227463667	498
Dickens[NNP]	0.0002227386521	499
resonance-chamber[NNS]	0.0002226351367	500
leading-tone[NN]	0.0002224820318	501
florid[JJ]	0.0001438729148	997
sound[VBZ]	0.000143856694	998
score[NNS]	0.0001437556948	999
rondo[NN]	0.0001435829476	1000
sweet[JJ]	0.0001435409753	1001
sense[NN]	0.0001434473773	1002
gesture[NNS]	9.09E-05	1997
hammer[NNS]	9.08E-05	1998
flow[NN]	9.08E-05	1999
sorrow[NN]	9.08E-05	2000
monophonic[JJ]	9.08E-05	2001
saint[NNS]	4.79E-05	4997
move[VBZ]	4.79E-05	4998
moderately[RB]	4.79E-05	4999
Cecilia[NNP]	4.79E-05	5000
Nibelung[NNP]	4.79E-05	5001
mean[VBD]	2.80E-05	9997
aloft[RB]	2.80E-05	9998
o'er[RB]	2.80E-05	9999
unaffected[JJ]	2.80E-05	10000
Stockhausen[NNP]	2.80E-05	10001
indulgent[JJ]	1.42E-05	19997
emulation[NN]	1.42E-05	19998
emerge[VB]	1.42E-05	19999
two-step[NNS]	1.42E-05	20000
Lauriett[NNP]	1.42E-05	20001
unfitness[NN]	5.86E-06	39997
Aryan[NNP]	5.86E-06	39998
Sirens[NNPS]	5.86E-06	39999
MACREADY[NNP]	5.86E-06	40000
fence[VBN]	5.85E-06	40001
offrir[FW]	3.18E-06	59997
postes[FW]	3.18E-06	59998
Dorf[NNP]	3.18E-06	59999
Dewing[NNP]	3.18E-06	60000
legitimise[VBN]	3.18E-06	60001

Validation

We compared Listening Experiences and the Reuters-21578 corpus [1] (used to benchmark news classification systems, does not include music as category).





- We matched the vector of each corpus with the music dictionary, and clustered the number of occurrences per score range (log scale in the pictures)
- We calculated a distribution score (sum(scores) / corpus vector length)
  - LE (vector length: 949301) is **0.000**114802266598619 74
  - Reuters-21578 (vector length: 1372059) is 4.6368513916777576e-05 (**0.0000**4636851)

The dictionary fits better LEs then Reuters

[1] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

Next step (Phase 2)

- Build a benchmark using the LEs and their original sources
- Design/generate a model by using the dictionary in combination with features of the LE to apply to incoming texts.
- Develop a system that generate annotations of texts and evaluate it on the benchmark.