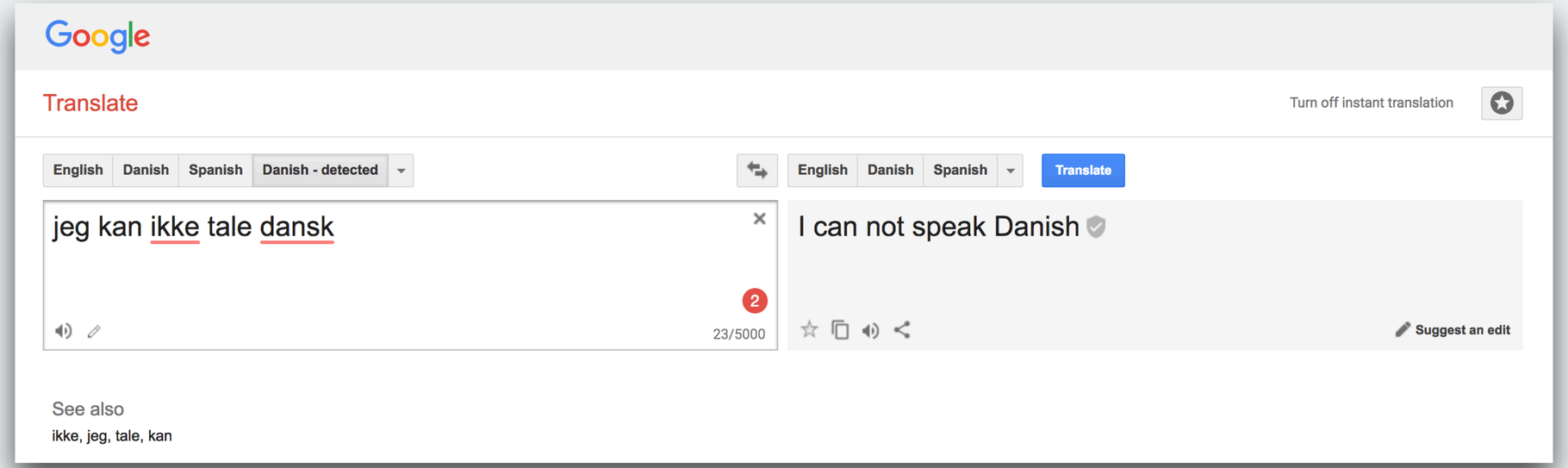
Computational Analysis of Big Data

Week 7

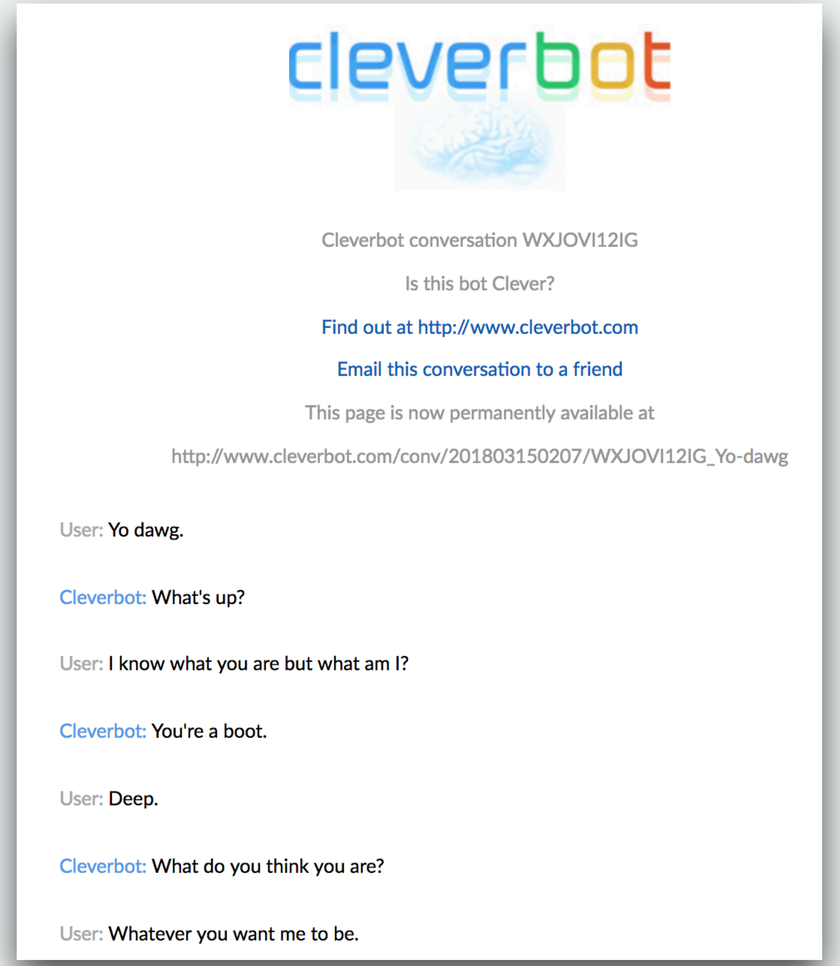
Natural Language Processing

NLP you know

Cross-language translation



Chat bots


cleverbot

Cleverbot conversation WXJOVI12IG

Is this bot Clever?

Find out at http://www.cleverbot.com

Email this conversation to a friend

This page is now permanently available at

http://www.cleverbot.com/conv/201803150207/WXJOVI12IG_Yo-dawg

User: Yo dawg.

Cleverbot: What's up?

User: I know what you are but what am I?

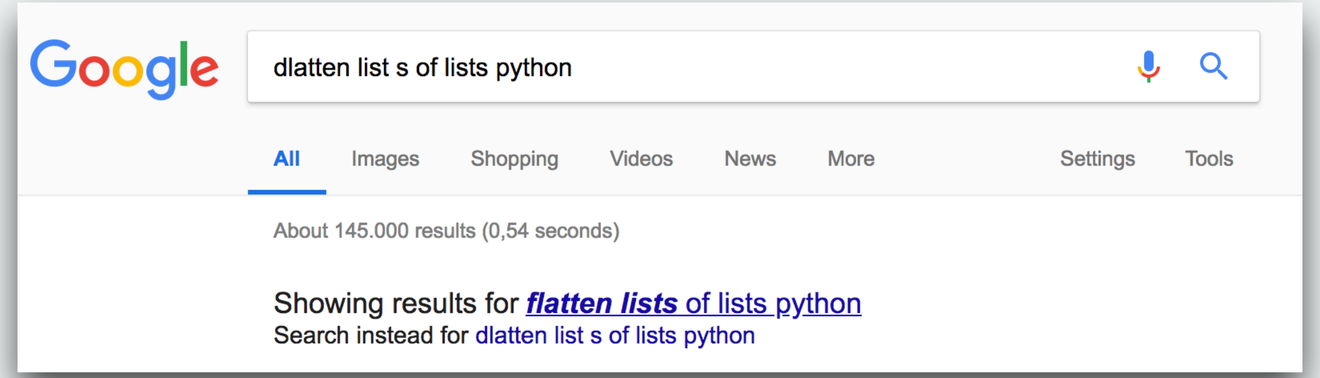
Cleverbot: You're a boot.

User: Deep.

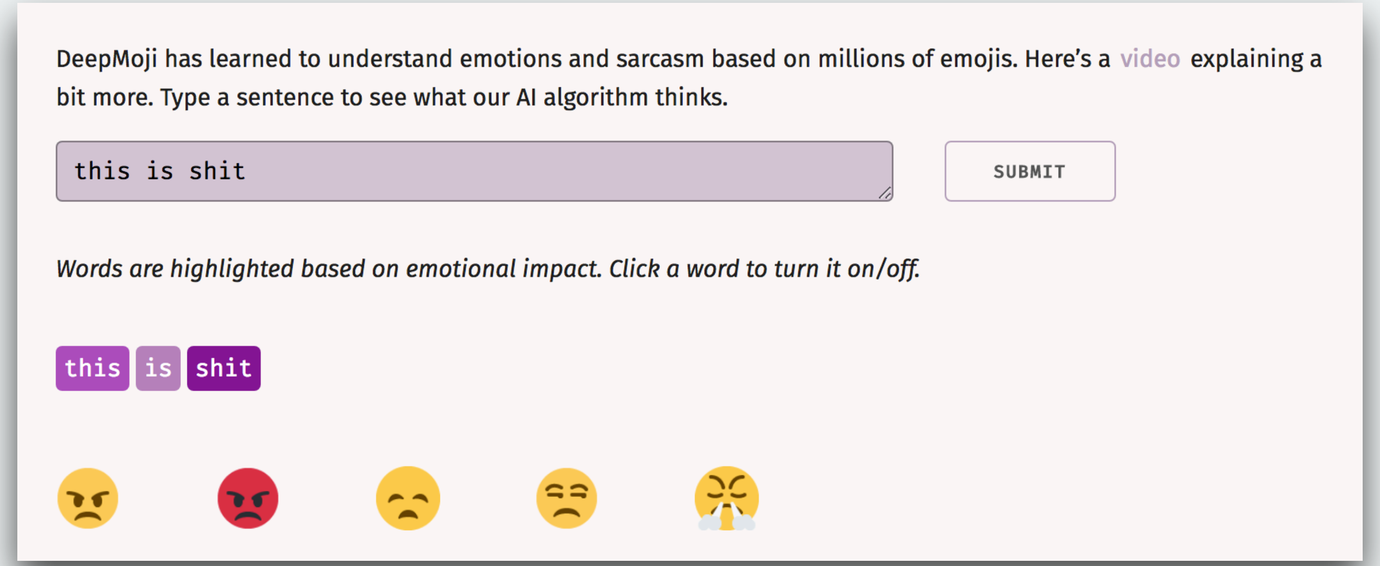
Cleverbot: What do you think you are?

User: Whatever you want me to be.

Word matching and autocorrect

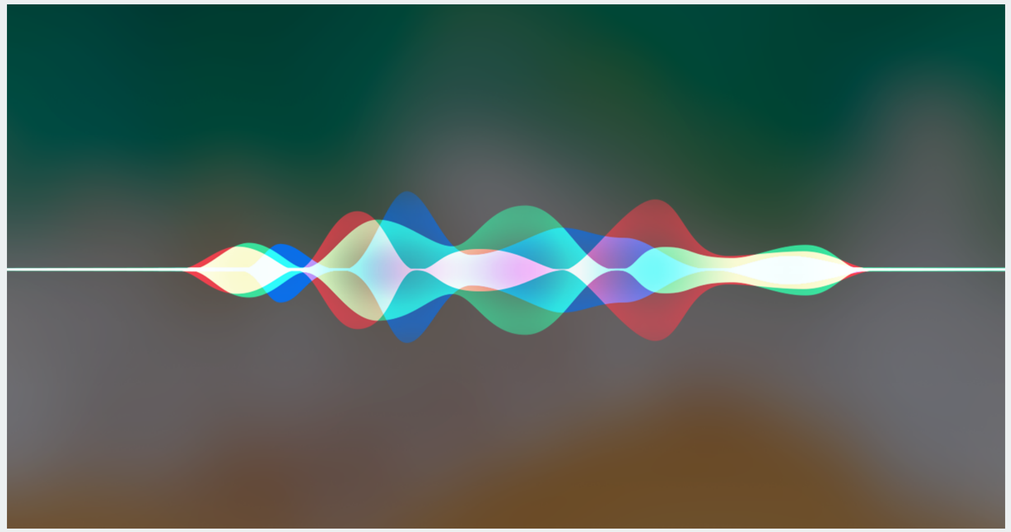


Emoji-detection

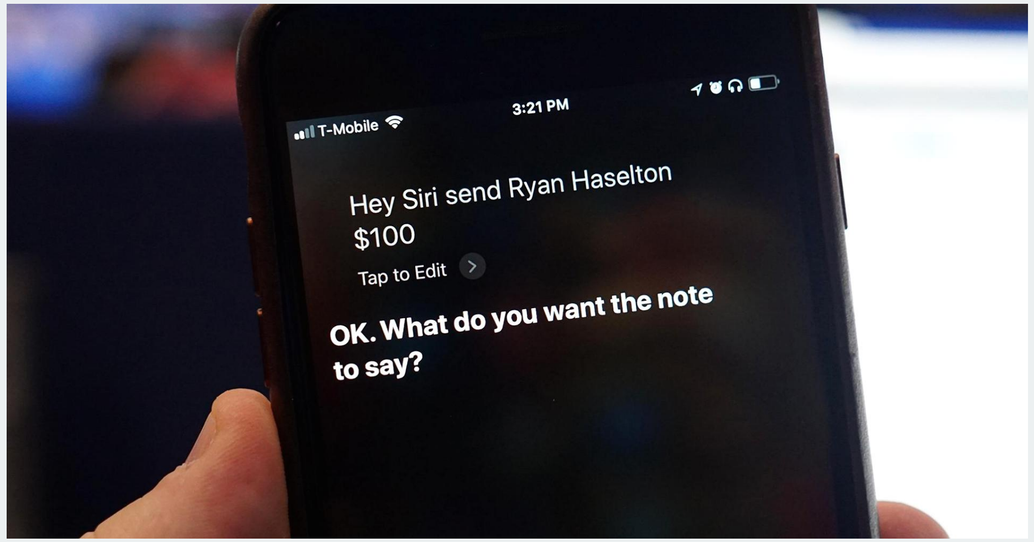


https://deepmoji.mit.edu/

Speech-to-text processing



to



Voice assistants

Input: Spoken words

Internally: Infer meaning and intent

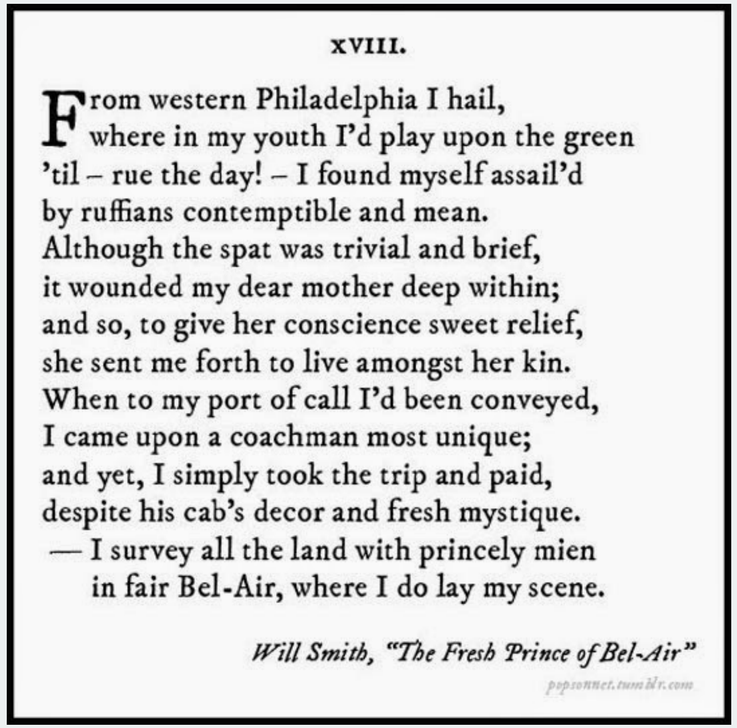
Output: Specific action

Speech to speech translation

https://www.youtube.com/watch?v=oQVQVt5H2QM

Text as data

Text as data



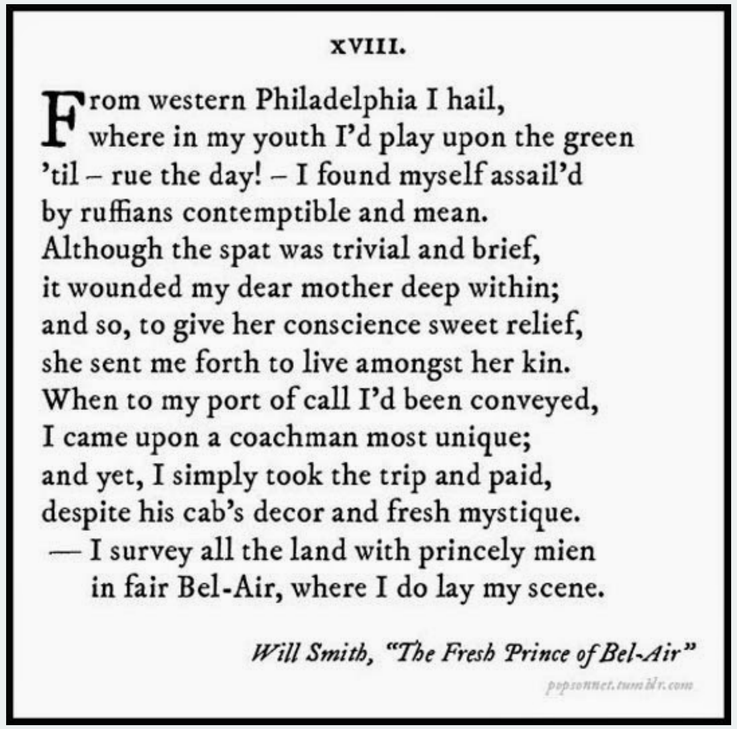
XVIII.

From western Philadelphia I hail,
where in my youth I'd play upon the green
'til – rue the day! – I found myself assail'd
by ruffians contemptible and mean.
Although the spat was trivial and brief,
it wounded my dear mother deep within;
and so, to give her conscience sweet relief,
she sent me forth to live amongst her kin.
When to my port of call I'd been conveyed,
I came upon a coachman most unique;
and yet, I simply took the trip and paid,
despite his cab's decor and fresh mystique.
— I survey all the land with princely mien
in fair Bel-Air, where I do lay my scene.

Will Smith, "The Fresh Prince of Bel-Air"

popsonnet.tumblr.com

Text as data



XVIII.

From western Philadelphia I hail,
where in my youth I'd play upon the green
'til – rue the day! – I found myself assail'd
by ruffians contemptible and mean.
Although the spat was trivial and brief,
it wounded my dear mother deep within;
and so, to give her conscience sweet relief,
she sent me forth to live amongst her kin.
When to my port of call I'd been conveyed,
I came upon a coachman most unique;
and yet, I simply took the trip and paid,
despite his cab's decor and fresh mystique.
— I survey all the land with princely mien
in fair Bel-Air, where I do lay my scene.

Will Smith, "The Fresh Prince of Bel-Air"

poponnet.tumblr.com

Extract markup


```
'From western Philadelphia I
hail,\nwhere in my youth
I\xe2\x80\x99d play upon the
green\n\xe2\x80\x99til
\xe2\x80\x94 rue the day!
\xe2\x80\x94 I found myself
assail\xe2\x80\x99d\nby ru{ans
contemptible and mean.
\nAlthough the spat was trivial
and brief,\nit wounded my dear
mother deep within;\nand so, to
give her conscience sweet
relief,\nshe sent me forth to
live amongst her kin.\nWhen to
my port of call I\xe2\x80\x99d
been convey\xe2\x80\x99d,\nI
came upon a coachman most
unique;\nand yet I simply took
the trip and paid,\ndespite his
cab\xe2\x80\x99s decor and
fresh mystique.\n\xc2\xa0
\xc2\xa0 \xc2\xa0\xe2\x80\x94 I
survey all the land with
princely mien\n\xc2\xa0
\xc2\xa0 \xc2\xa0in fair Bel-
Air, where I do lay my scene.
\n\nWill Smith, \xe2\x80\x9cThe
Fresh Prince of Bel-
Air\xe2\x80\x9d'
```

The sequential approach

```
'From western Philadelphia I
hail,\nwhere in my youth
I\xe2\x80\x99d play upon the
green\n\xe2\x80\x99til
\xe2\x80\x94 rue the day!
\xe2\x80\x94 I found myself
assail\xe2\x80\x99d\nby ru{ans
contemptible and mean.
\nAlthough the spat was trivial
and brief,\nit wounded my dear
mother deep within;\nand so, to
give her conscience sweet
relief,\nshe sent me forth to
live amongst her kin.\nWhen to
my port of call I\xe2\x80\x99d
been convey\xe2\x80\x99d,\nI
came upon a coachman most
unique;\nand yet I simply took
the trip and paid,\ndespite his
cab\xe2\x80\x99s decor and
fresh mystique.\n\xc2\xa0
\xc2\xa0 \xc2\xa0\xe2\x80\x94 I
survey all the land with
princely mien\n\xc2\xa0
\xc2\xa0 \xc2\xa0in fair Bel-
Air, where I do lay my scene.
\n\nWill Smith, \xe2\x80\x9cThe
Fresh Prince of Bel-
Air\xe2\x80\x9d'
```

Character-level encoding →

```
'F': [0 0 0 0 0 1 … 0 0 0]
'r': [0 0 0 0 0 0 … 0 0 0]
'o': [0 0 0 0 0 0 … 0 0 0]
'm': [0 0 0 0 0 0 … 0 0 0]
' ': [0 0 0 0 0 0 … 1 0 0]
'w': [0 0 0 0 0 0 … 0 0 0]
'e': [0 0 0 0 1 0 … 0 0 0]
                …
\x80: [0 0 0 0 0 0 … 0 1 0]
\x9d: [0 0 0 0 0 0 … 0 0 1]
```

One-hot encoding: Represent things as a vector of 0s and a single 1

The sequential approach

```
'From western Philadelphia I
hail,\nwhere in my youth
I\xe2\x80\x99d play upon the
green\n\xe2\x80\x99til
\xe2\x80\x94 rue the day!
\xe2\x80\x94 I found myself
assail\xe2\x80\x99d\nby ru{ans
contemptible and mean.
\nAlthough the spat was trivial
and brief,\nit wounded my dear
mother deep within;\nand so, to
give her conscience sweet
relief,\nshe sent me forth to
live amongst her kin.\nWhen to
my port of call I\xe2\x80\x99d
been convey\xe2\x80\x99d,\nI
came upon a coachman most
unique;\nand yet I simply took
the trip and paid,\ndespite his
cab\xe2\x80\x99s decor and
fresh mystique.\n\xc2\xa0
\xc2\xa0 \xc2\xa0\xe2\x80\x94 I
survey all the land with
princely mien\n\xc2\xa0
\xc2\xa0 \xc2\xa0in fair Bel-
Air, where I do lay my scene.
\n\nWill Smith, \xe2\x80\x9cThe
Fresh Prince of Bel-
Air\xe2\x80\x9d'
```

Character-level encoding →

```
'F': [0 0 0 0 0 1 … 0 0 0]
'r': [0 0 0 0 0 0 … 0 0 0]
'o': [0 0 0 0 0 0 … 0 0 0]
'm': [0 0 0 0 0 0 … 0 0 0]
' ': [0 0 0 0 0 0 … 1 0 0]
'w': [0 0 0 0 0 0 … 0 0 0]
'e': [0 0 0 0 1 0 … 0 0 0]
                …
\x80: [0 0 0 0 0 0 … 0 1 0]
\x9d: [0 0 0 0 0 0 … 0 0 1]
```

Word-level encoding →

```
'From': [1 0 0 0 0 0 … 0 0 0]
'western': [0 1 0 0 0 0 … 0 0 0]
'Philadelphia': [0 0 1 0 0 0 … 0 0 0]
'I': [0 0 0 1 0 0 … 0 0 0]
'hail': [0 0 0 0 1 0 … 0 0 0]
'where': [0 0 0 0 0 1 … 0 0 0]
'in': [0 0 0 0 0 0 … 0 0 0]
                …
'of': [0 0 0 0 0 0 … 0 0 0]
'Bel-Air': [0 0 0 0 0 0 … 0 0 1]
```

One-hot encoding: Represent things as a vector of 0s and a single 1

The aggregate approach

```
'From western Philadelphia I
hail,\nwhere in my youth
I\xe2\x80\x99d play upon the
green\n\xe2\x80\x99til
\xe2\x80\x94 rue the day!
\xe2\x80\x94 I found myself
assail\xe2\x80\x99d\nby ru{ans
contemptible and mean.
\nAlthough the spat was trivial
and brief,\nit wounded my dear
mother deep within;\nand so, to
give her conscience sweet
relief,\nshe sent me forth to
live amongst her kin.\nWhen to
my port of call I\xe2\x80\x99d
been convey\xe2\x80\x99d,\nI
came upon a coachman most
unique;\nand yet I simply took
the trip and paid,\ndespite his
cab\xe2\x80\x99s decor and
fresh mystique.\n\xc2\xa0
\xc2\xa0 \xc2\xa0\xe2\x80\x94 I
survey all the land with
princely mien\n\xc2\xa0
\xc2\xa0 \xc2\xa0in fair Bel-
Air, where I do lay my scene.
\n\nWill Smith, \xe2\x80\x9cThe
Fresh Prince of Bel-
Air\xe2\x80\x9d'
```

“Bag of Symbols” →

	a	b	c	d	e	f	…	\x20	\x80	\x9d
whole poem	[31	5	14	22	60	9	…	106	11	1]

“Bag of Words” →

	From	western	Philadelphia	I	hail	where	…	Prince	of	Bel-Air
whole poem	[1	1	1	8	1	2	…	1	2	2]

“Bag” encoding: Count number of occurrences of each element (similar to histogram)

The aggregate approach

```
'From western Philadelphia I
hail,\nwhere in my youth
I\xe2\x80\x99d play upon the
green\n\xe2\x80\x99til
\xe2\x80\x94 rue the day!
\xe2\x80\x94 I found myself
assail\xe2\x80\x99d\nby ru{ans
contemptible and mean.
\nAlthough the spat was trivial
and brief,\nit wounded my dear
mother deep within;\nand so, to
give her conscience sweet
relief,\nshe sent me forth to
live amongst her kin.\nWhen to
my port of call I\xe2\x80\x99d
been convey\xe2\x80\x99d,\nI
came upon a coachman most
unique;\nand yet I simply took
the trip and paid,\ndespite his
cab\xe2\x80\x99s decor and
fresh mystique.\n\xc2\xa0
\xc2\xa0 \xc2\xa0\xe2\x80\x94 I
survey all the land with
princely mien\n\xc2\xa0
\xc2\xa0 \xc2\xa0in fair Bel-
Air, where I do lay my scene.
\n\nWill Smith, \xe2\x80\x9cThe
Fresh Prince of Bel-
Air\xe2\x80\x9d'
```

“Bag of Symbols” →

	a	b	c	d	e	f	…	\x20	\x80	\x9d
whole poem	31	5	14	22	60	9	…	106	11	1
another poem	34	5	84	13	50	1	…	431	10	2
yet another poem	22	1	12	19	12	5	…	123	19	1
…										
the last of many poems	19	3	13	77	13	8	…	213	43	4

“Bag of Words” →

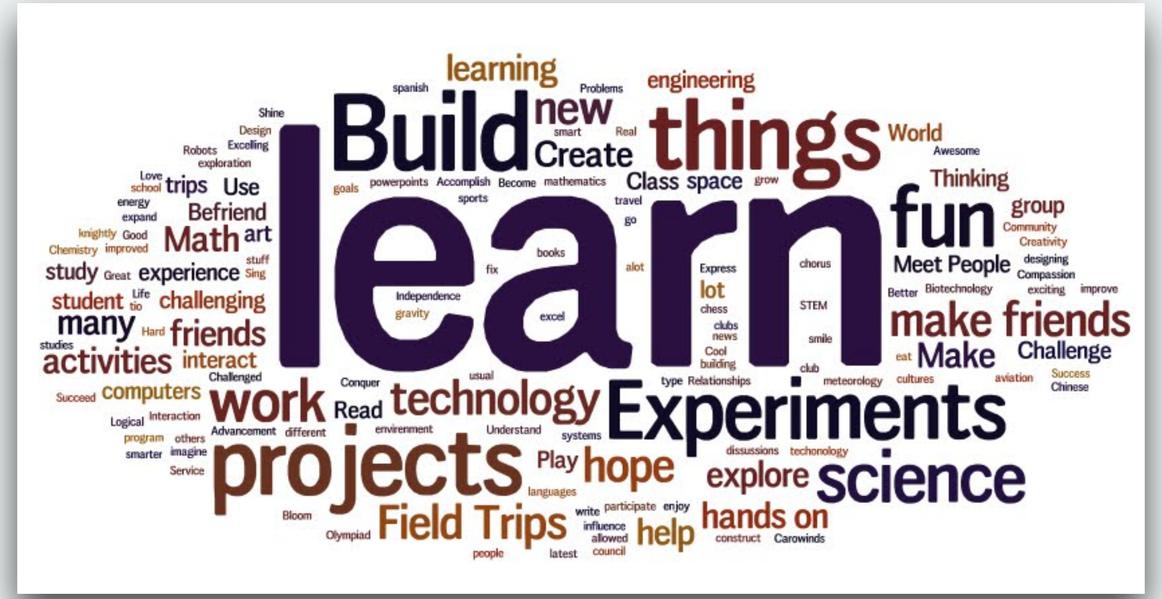
	From	western	Philadelphia	I	hail	where	…	Prince	of	Bel-Air
whole poem	1	1	1	8	1	2	…	1	2	2
another poem	4	0	0	9	0	6	…	0	9	0
yet another poem	9	0	3	7	2	2	…	0	0	0
…										
the last of many poems	2	7	0	1	0	0	…	0	4	0

“Corpus”: A collection of documents represented as a 2d array (list of vectors)

“Bag” encoding: Count number of occurrences of each element (similar to histogram)

Analysis methods

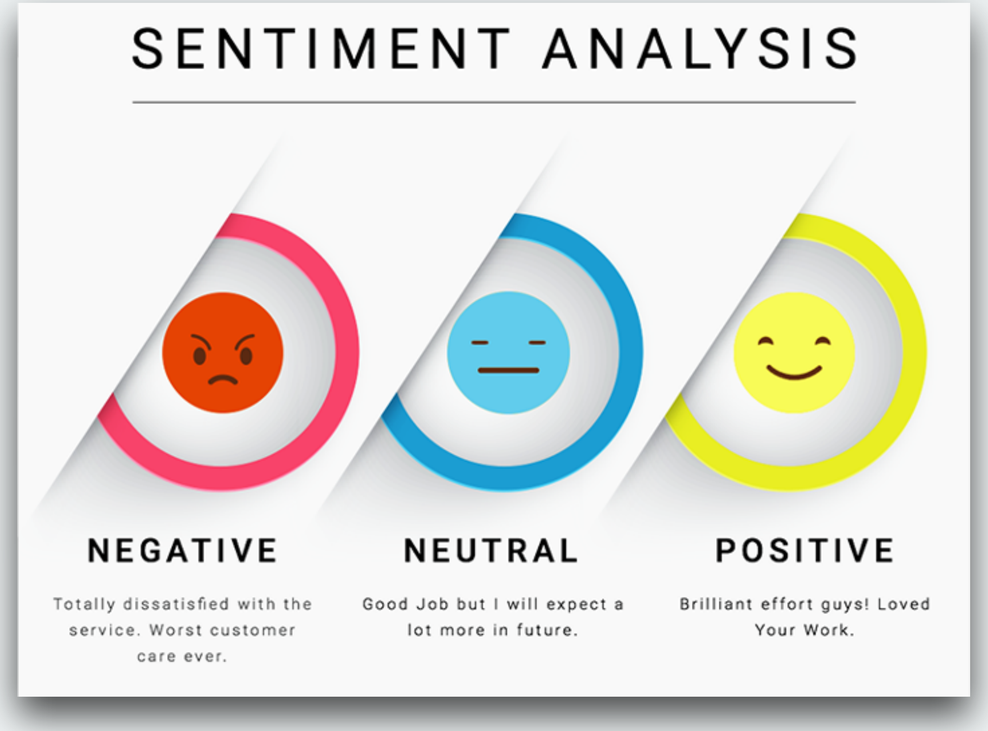
Word clouds



Algorithm:

1. Have a document (poem, book, article, …)
2. Count frequency of words (e.g. using BoW)
3. Print the words with sizes relative to their frequency
4. Put larger words in center and smaller words in periphery
5. Impress shareholders that think computers are magic

Sentiment analysis



Algorithm:

1. Have a document (poem, book, article, …)
2. Count frequency of words (e.g. using BoW)
3. Map each word to a pre-estimated "sentiment score”
4. Take frequency weighted average of sentiment scores for all words
5. Measure if text is negative or positive

Term Frequency - Inverse Document Frequency

	From	western	Philadelphia	I	hail	where	:	Prince	of	Bel-Air
whole poem	1	1	1	8	1	2	…	1	2	2
another poem	4	0	0	9	0	6	…	0	9	0
yet another poem	9	0	3	7	2	2	…	0	0	0
…										
the last of many poems	2	7	0	1	0	0	…	0	4	0

whole poem	0	0	0.1	0	0	0	…	0.2	0	0.7
another poem	0.3	0	0.3	0	0	0.2	…	0	0.2	0
yet another poem	0.5	0	0.2	0.1	0.2	0	…	0	0	0
…										
the last of many poems	0	0.7	0	0	0	0	…	0	0.3	0

Algorithm:

1. Have BoW representation of document
2. *TF-step:* Normalize word frequency in each document (so rows sum to 1)
3. *IDF-step:* Estimate the document frequency of each word (in what fraction of document e.g. ‘western' occurs), and divide each column element with this value.
4. You now have a matrix, where a (document, word) index value explains how important a given word is to that document

Other tasks

Syntax:

1. Grammar induction
2. Part-of-speech tagging
3. Lemmatization/stemming
4. Sentence breaking
5. Word segmentation

Semantics:

1. Machine translation
2. Natural language generation
3. Question answering
4. Topic modeling
5. Word sense disambiguation
6. Automatic summarization

Speech:

1. Speech recognition
2. Speech segmentation
3. Text-to-speech
4. Speech-to-speech