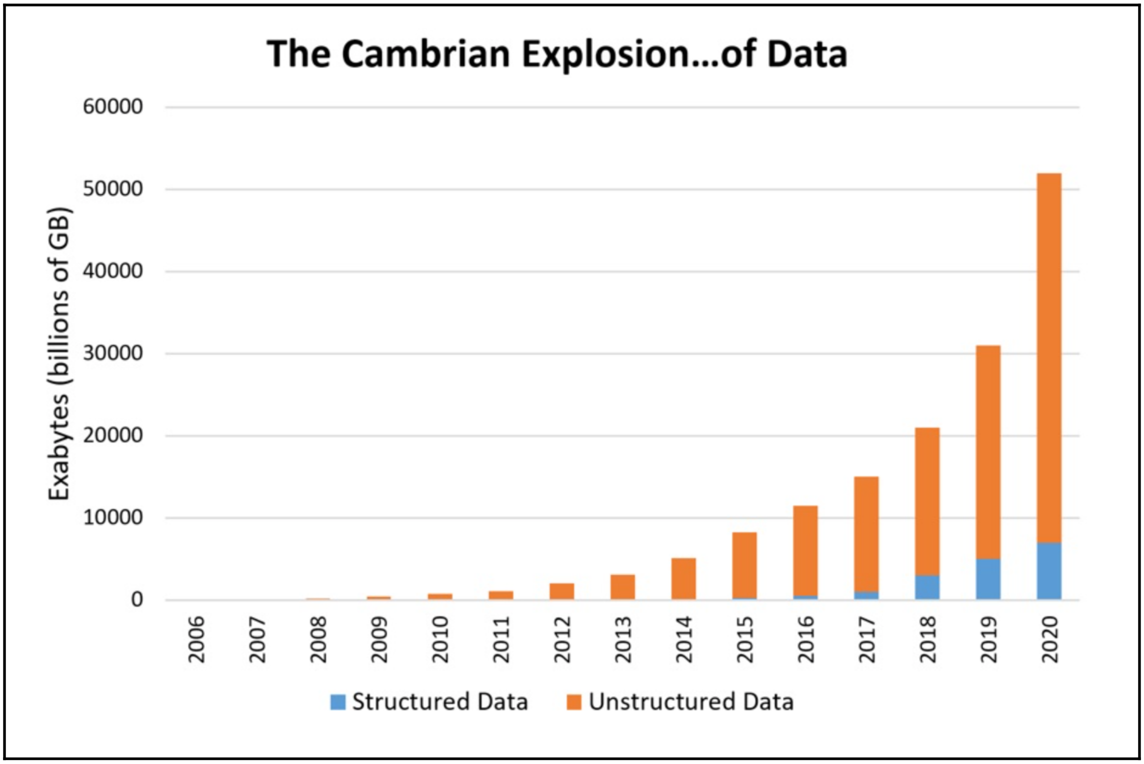
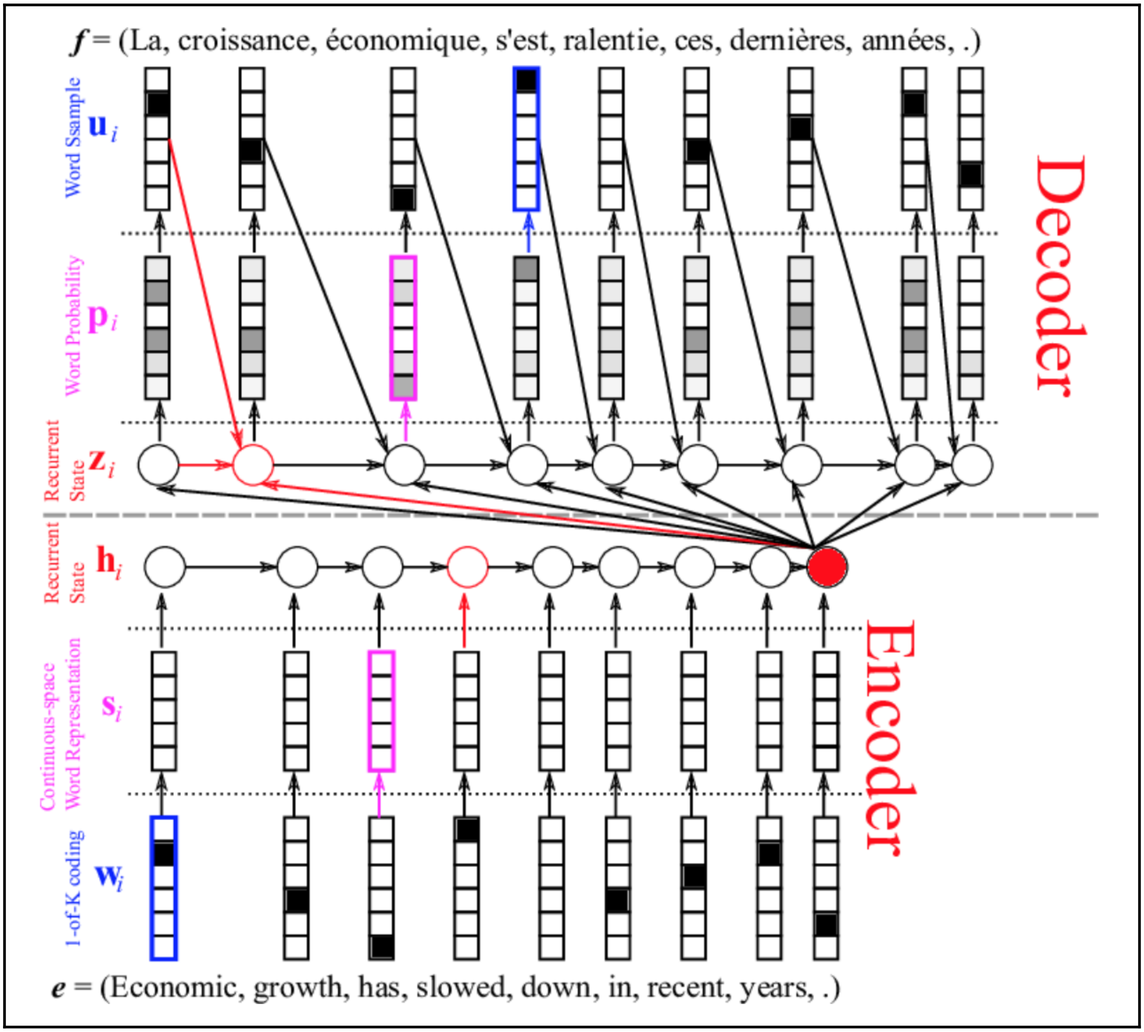
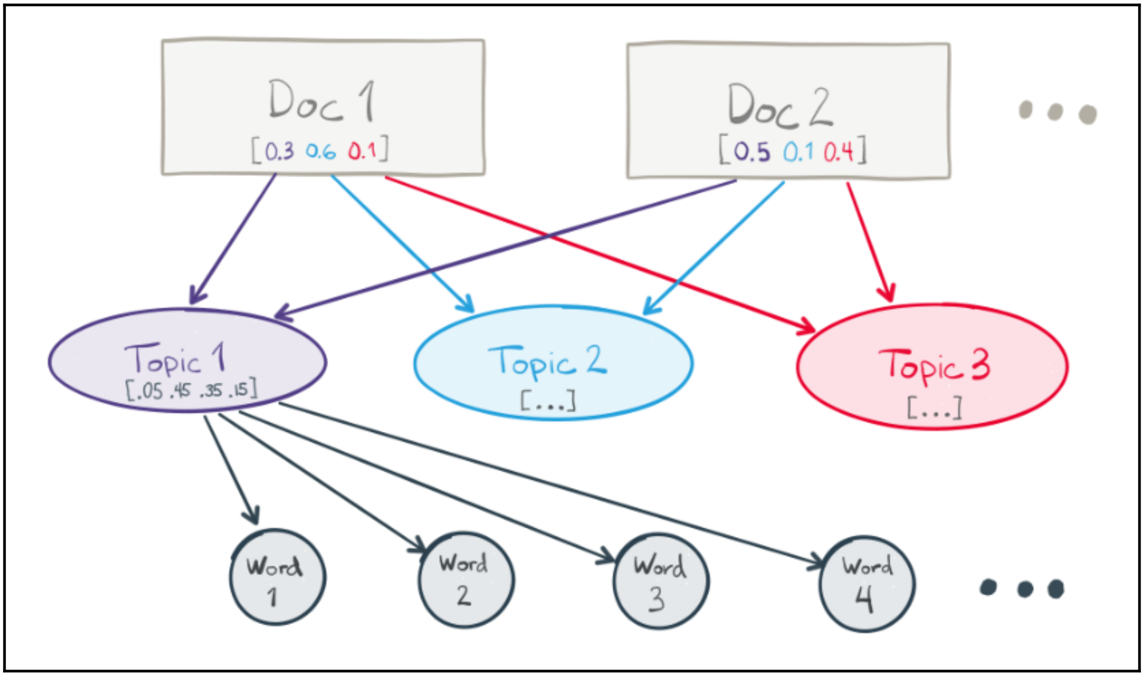
Chapter 1: What is Text Analysis?





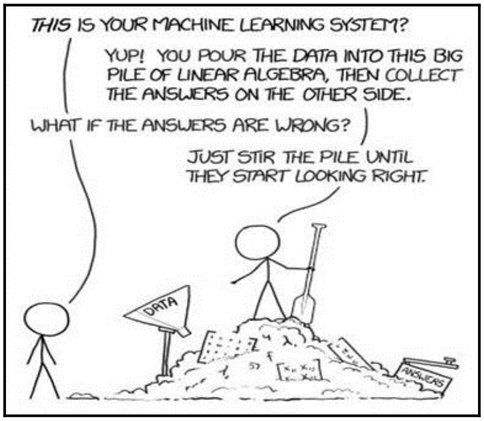
f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
$\mathbf{u}_i$
Word Probability
$\mathbf{p}_i$
Recurrent State
$\mathbf{z}_i$
Decoder
Recurrent State
$\mathbf{h}_i$
Continuous-space Word Representation
$\mathbf{s}_i$
1-of-K coding
$\mathbf{w}_i$
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)


Doc 1
[0.3 0.6 0.1]
Doc 2
[0.5 0.1 0.4]
Topic 1
[.05 .45 .35 .15]
Topic 2
[...]
Topic 3
[...]
Word 1
Word 2
Word 3
Word 4

Dataset Name	Brief description	Preprocessing	Instances	Format	Default Task	Created (updated)	Reference	Creator
Amazon reviews	US product reviews from Amazon.com.	None.	~ 82M	Text	Classification, sentiment analysis	2015	[131]	McAuley et al.
OpinRank Review Dataset	Reviews of cars and hotels from Edmunds.com and TripAdvisor respectively.	None.	42,230 / ~259,000 respectively	Text	Sentiment analysis, clustering	2011	[132][133]	K. Ganesan et al.
MovieLens	22,000,000 ratings and 580,000 tags applied to 33,000 movies by 240,000 users.	None.	~ 22M	Text	Regression, clustering, classification	2016	[134]	GroupLens Research
Yahoo! Music User Ratings of Musical Artists	Over 10M ratings of artists by Yahoo users.	None described.	~ 10M	Text	Clustering, regression	2004	[135][136]	Yahoo!
Car Evaluation Data Set	Car properties and their overall acceptability.	Six categorical features given.	1728	Text	Classification	1997	[137][138]	M. Bohanec
YouTube Comedy Slam Preference Dataset	User vote data for pairs of videos shown on YouTube. Users voted on funnier videos.	Video metadata given.	1,138,562	Text	Classification	2012	[139][140]	Google
Skytrax User Reviews Dataset	User reviews of airlines, airports, seats, and lounges from Skytrax.	Ratings are fine-grain and include many aspects of airport experience.	41396	Text	Classification, regression	2015	[141]	Q. Nguyen
Teaching Assistant Evaluation Dataset	Teaching assistant reviews.	Features of each instance such as class, class size, and instructor are given.	151	Text	Classification	1997	[142][143]	W. Loh et al.

Twitter and tweets [edit]

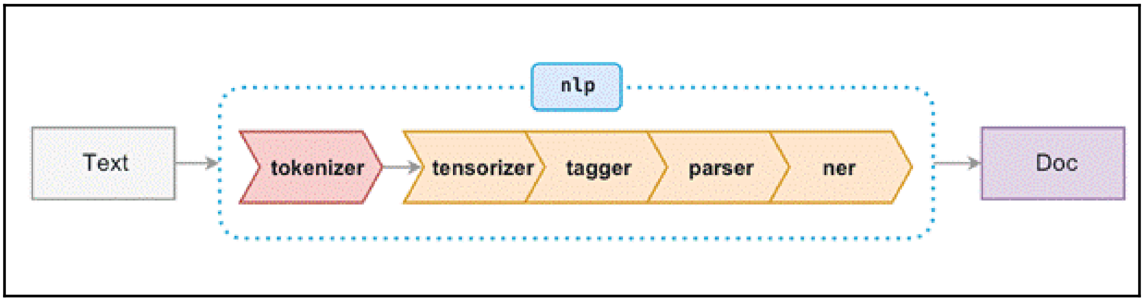
Dataset Name	Brief description	Preprocessing	Instances	Format	Default Task	Created (updated)	Reference	Creator
Sentiment140	Tweet data from 2009 including original text, time stamp, user and sentiment.	Classified using distant supervision from presence of emoticon in tweet.	1,578,627	Tweets, comma, separated values	Sentiment analysis	2009	[161][162]	A. Go et al.
ASU Twitter Dataset	Twitter network data, not actual tweets. Shows connections between a large number of users.	None.	11,316,811 users, 85,331,846 connections	Text	Clustering, graph analysis	2009	[163][164]	R. Zafarani et al.
SNAP Social Circles: Twitter Database	Large twitter network data.	Node features, circles, and ego networks.	1,768,149	Text	Clustering, graph analysis	2012	[165][166]	J. McAuley et al.
Twitter Dataset for Arabic Sentiment Analysis	Arabic tweets.	Samples hand-labeled as positive or negative.	2000	Text	Classification	2014	[167][168]	N. Abdulla
Buzz in Social Media Dataset	Data from Twitter and Tom's Hardware. This dataset focuses on specific buzz topics being discussed on those sites.	Data is windowed so that the user can attempt to predict the events leading up to social media buzz.	140,000	Text	Regression, Classification	2013	[169][170]	F. Kawala et al.
Paraphrase and Semantic Similarity in Twitter (PIT)	This dataset focuses on whether tweets have (almost) same meaning/information or not. Manually labeled.	tokenization, part-of-speech and named entity tagging	18,762	Text	Regression, Classification	2015	[171][172]	Xu et al.

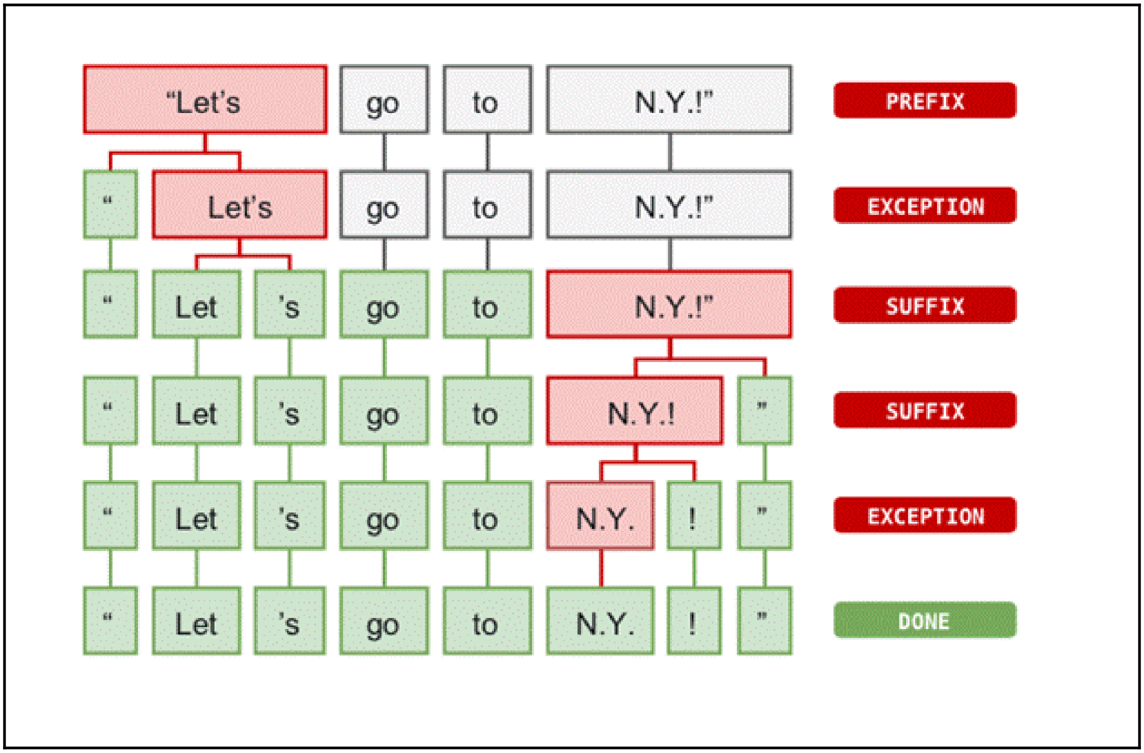


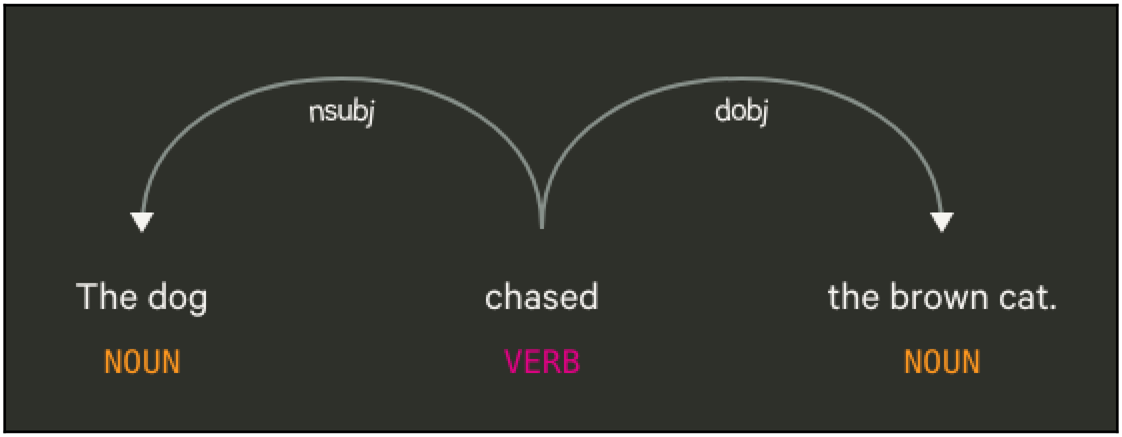
Feature comparison

Here's a quick comparison of the functionalities offered by spaCy, SyntaxNet, NLTK and CoreNLP.

	SPACY	SYNTAXNET	NLTK	CORENLP
Programming language	Python	C++	Python	Java
Neural network models	🟢	🟢	🔴	🟢
Integrated word vectors	🟢	🔴	🔴	🔴
Multi-language support	🟢	🟢	🟢	🟢
Tokenization	🟢	🟢	🟢	🟢
Part-of-speech tagging	🟢	🟢	🟢	🟢
Sentence segmentation	🟢	🟢	🟢	🟢
Dependency parsing	🟢	🟢	🔴	🟢
Entity recognition	🟢	🔴	🟢	🟢
Coreference resolution	🔴	🔴	🔴	🟢

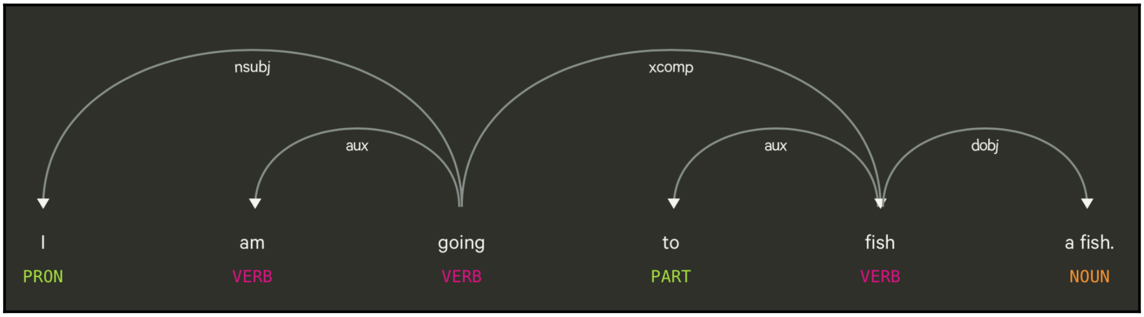


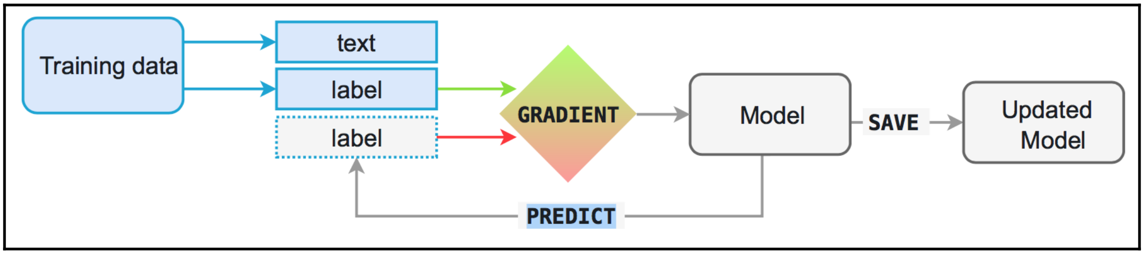

"Let's
go
to
N.Y.!"
PREFIX
"
Let's
EXCEPTION
Let
's
SUFFIX
N.Y.!
"
SUFFIX
N.Y.
!
EXCEPTION
DONE

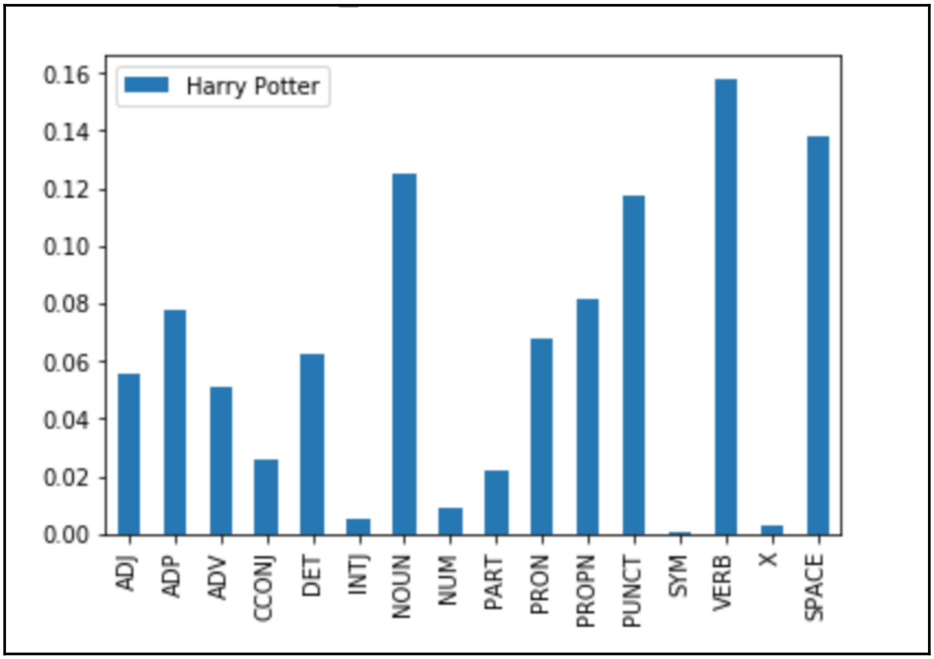

nsubj
dobj
The dog
chased
the brown cat.
NOUN
VERB
NOUN

Chapter 5: POS-Tagging and Its Applications

POS	DESCRIPTION	EXAMPLES
`ADJ`	adjective	*big, old, green, incomprehensible, first*
`ADP`	adposition	*in, to, during*
`ADV`	adverb	*very, tomorrow, down, where, there*
`AUX`	auxiliary	*is, has (done), will (do), should (do)*
`CONJ`	conjunction	*and, or, but*
`CCONJ`	coordinating conjunction	*and, or, but*
`DET`	determiner	*a, an, the*
`INTJ`	interjection	*psst, ouch, bravo, hello*
`NOUN`	noun	*girl, cat, tree, air, beauty*
`NUM`	numeral	*1, 2017, one, seventy-seven, IV, MMXIV*
`PART`	particle	*'s, not,*
`PRON`	pronoun	*I, you, he, she, myself, themselves, somebody*
`PROPN`	proper noun	*Mary, John, Londin, NATO, HBO*
`PUNCT`	punctuation	*., (,), ?*
`SCONJ`	subordinating conjunction	*if, while, that*
`SYM`	symbol	*$, %, §, ©, +, −, ×, ÷, =, :), 😝*
`VERB`	verb	*run, runs, running, eat, ate, eating*
`X`	other	*sfpksdpsxmsa*
`SPACE`	space	


nsubj
aux
xcomp
aux
dobj
I
PRON
am
VERB
going
VERB
to
PART
fish
VERB
a fish.
NOUN


Training data
text
label
label
GRADIENT
Model
SAVE
Updated Model
PREDICT


Harry Potter
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
ADJ
ADP
ADV
CCONJ
DET
INTJ
NOUN
NUM
PART
PRON
PROPN
PUNCT
SYM
VERB
X
SPACE

TYPE	DESCRIPTION
`PER`	Named person or family.
`LOC`	Name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains).
`ORG`	Named corporate, governmental, or other organizational entity.
`MISC`	Miscellaneous entities, e.g. events, nationalities, products or works of art.

TYPE	DESCRIPTION
`PERSON`	People, including fictional.
`NORP`	Nationalities or religious or political groups.
`FACILITY`	Buildings, airports, highways, bridges, etc.
`ORG`	Companies, agencies, institutions, etc.
`GPE`	Countries, cities, states.
`LOC`	Non-GPE locations, mountain ranges, bodies of water.
`PRODUCT`	Objects, vehicles, foods, etc. (Not services.)
`EVENT`	Named hurricanes, battles, wars, sports events, etc.
`WORK_OF_ART`	Titles of books, songs, etc.
`LAW`	Named documents made into laws.
`LANGUAGE`	Any named language.
`DATE`	Absolute or relative dates or periods.
`TIME`	Times smaller than a day.
`PERCENT`	Percentage, including "%".
`MONEY`	Monetary values, including unit.
`QUANTITY`	Measurements, as of weight or distance.
`ORDINAL`	"first", "second", etc.
`CARDINAL`	Numerals that do not fall under another type.

TAG	DESCRIPTION
`B EGIN`	The first token of a multi-token entity.
`I N`	An inner token of a multi-token entity.
`L AST`	The final token of a multi-token entity.
`U NIT`	A single-token entity.
`O UT`	A non-entity token.

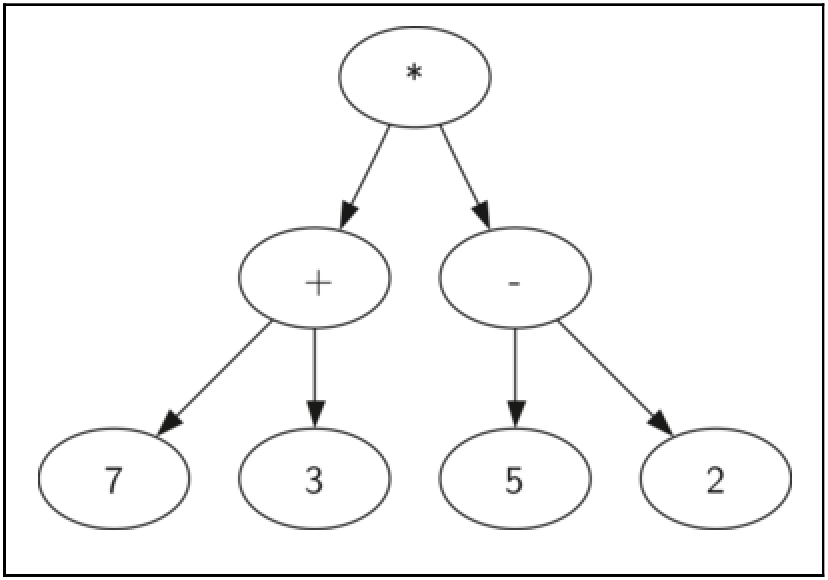
Elon Musk PERSON apparently wasn't aware that his company SpaceX had a Facebook ORG page. The SpaceX and Tesla PRODUCT CEO has responded to a comment on Twitter GPE calling for him to take down the SpaceX, Tesla and Elon Musk ORG official pages in support of the #deletefacebook movement by first ORDINAL acknowledging he didn't know one existed, and then following up with promises that he would indeed take them down.

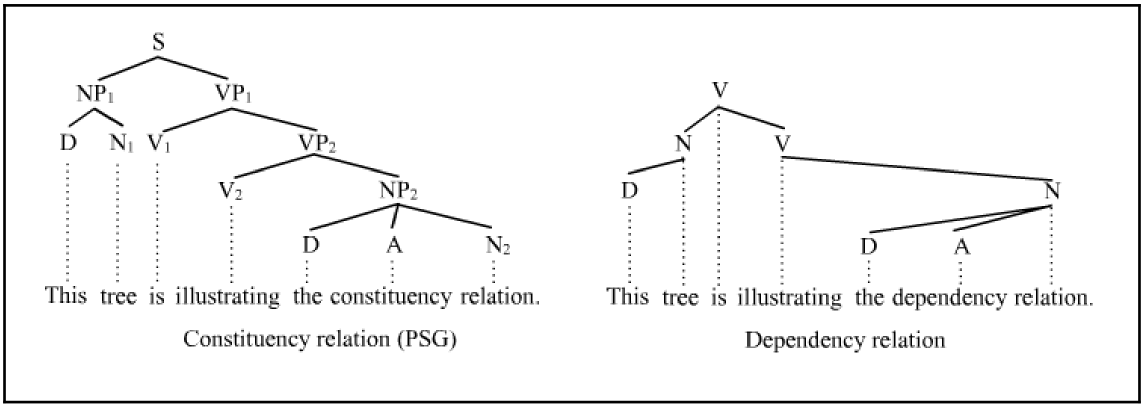
He's done just that, as the SpaceX NORP Facebook page is now gone, after having been live earlier today DATE (as you can see from the screenshot included taken at around 12:10 PM ET) TIME .

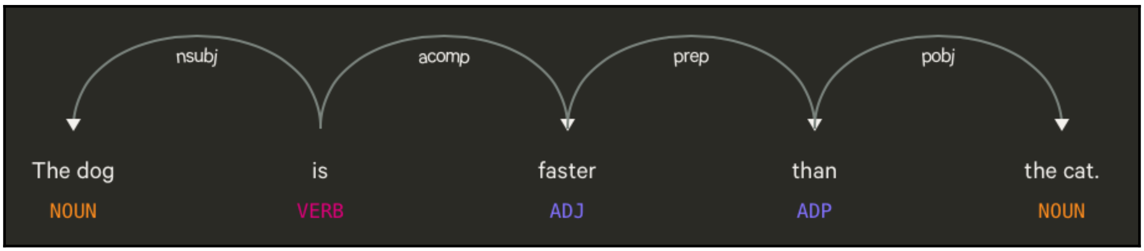
Emmanuel Jean-Michel Frédéric Macron PERSON is a French NORP politician serving as President of France GPE and ex officio Co-Prince of Andorra LOC since 14 May 2017 DATE .

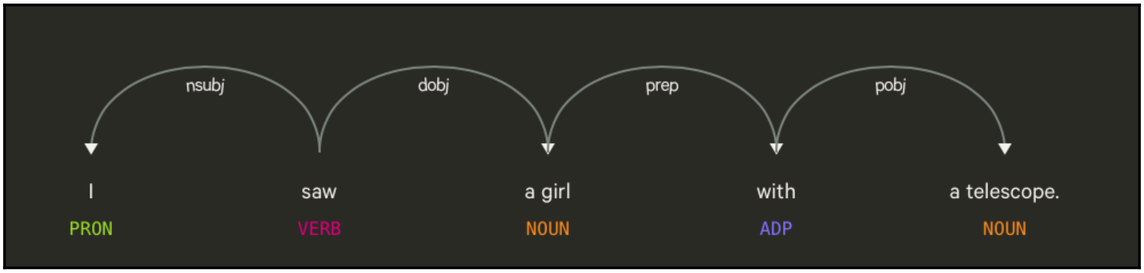
Before entering politics, he was a senior civil servant and investment banker. He studied philosophy at Paris Nanterre University ORG , completed a Master's of Public Affairs ORG at Sciences Po, and graduated from the École nationale d'administration (PRODUCT ÉNA ORG) in 2004 DATE . He worked at the Inspectorate General of Finances ORG , and later became an investment banker at Rothschild & Cie Banque ORG .

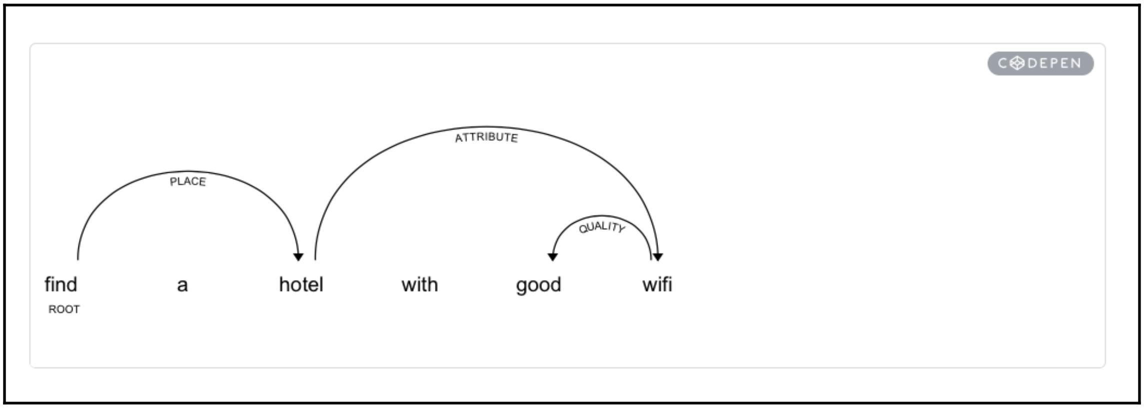
Chapter 7: Dependency Parsing



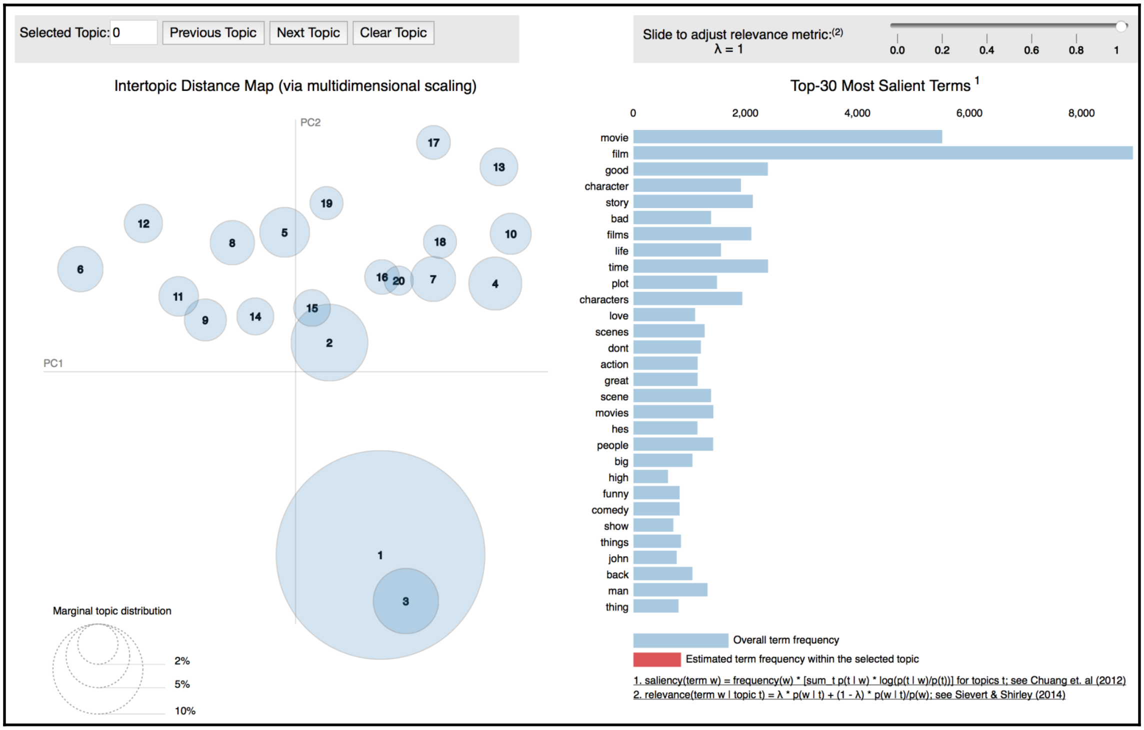


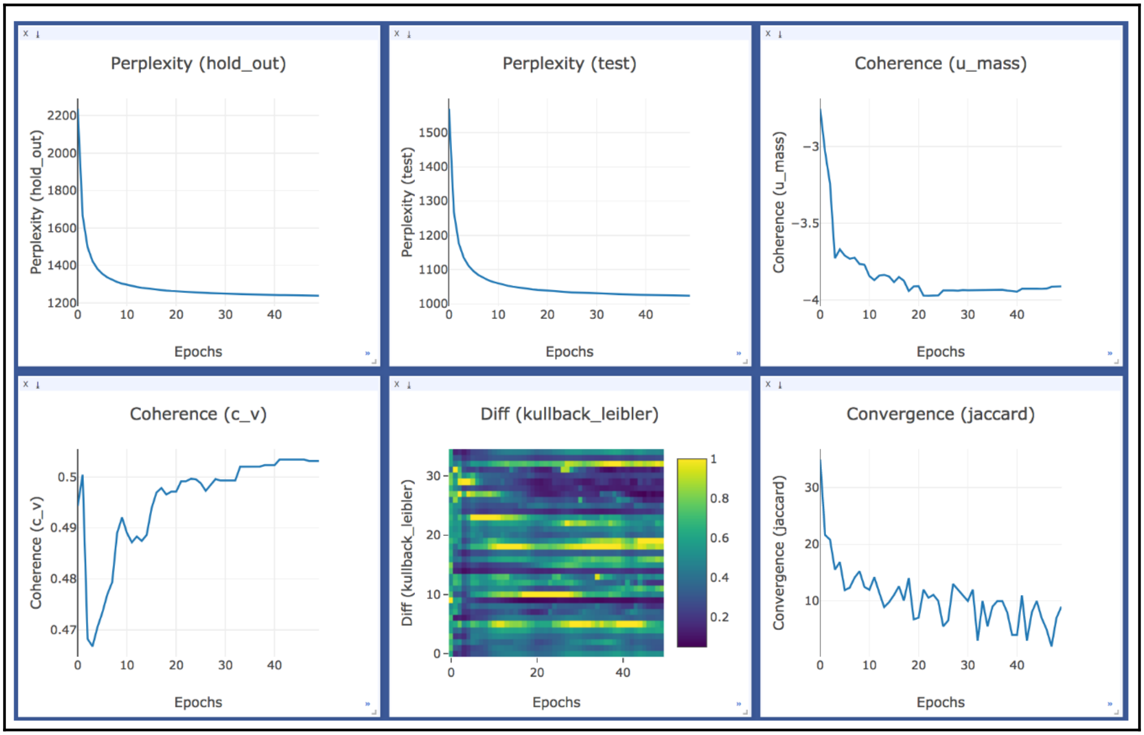


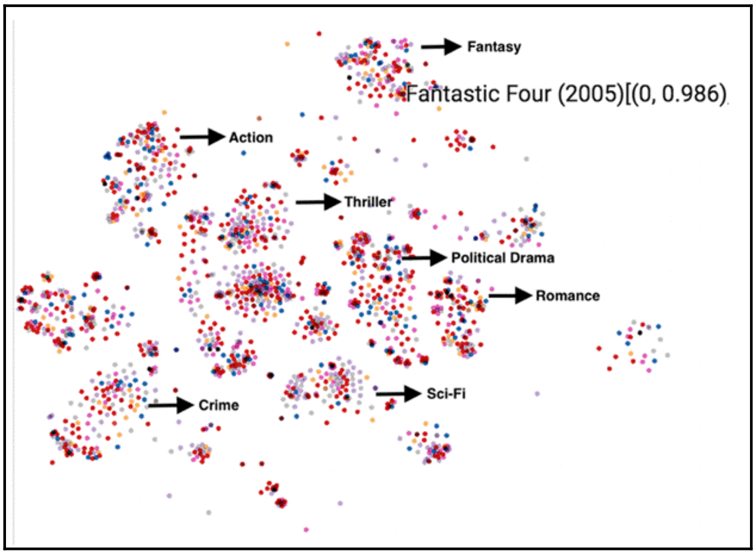

nsubj
dobj
prep
pobj
I
saw
a girl
with
a telescope.
PRON
VERB
NOUN
ADP
NOUN

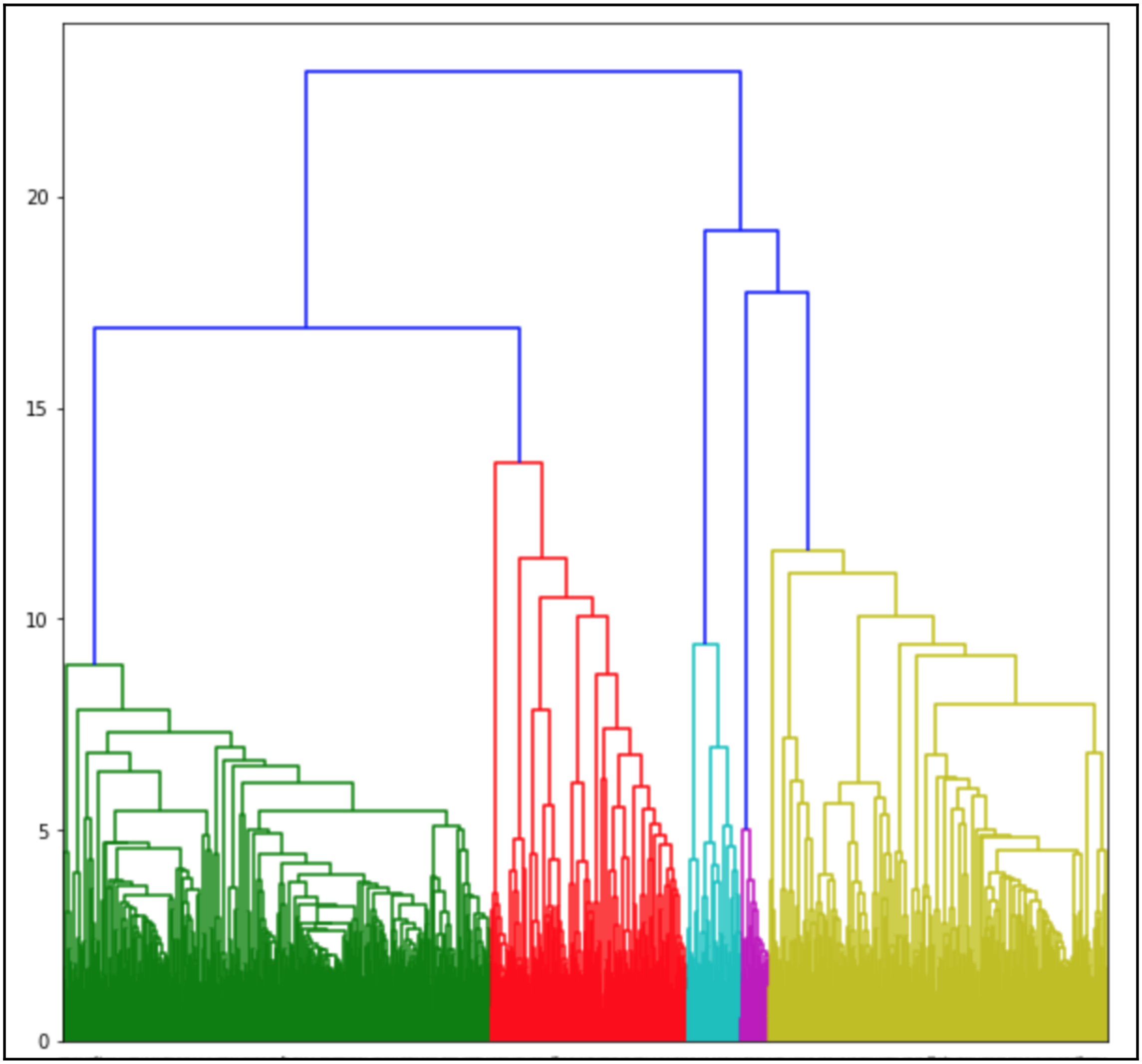

CODEPEN
ATTRIBUTE
PLACE
QUALITY
find
ROOT
a
hotel
with
good
wifi

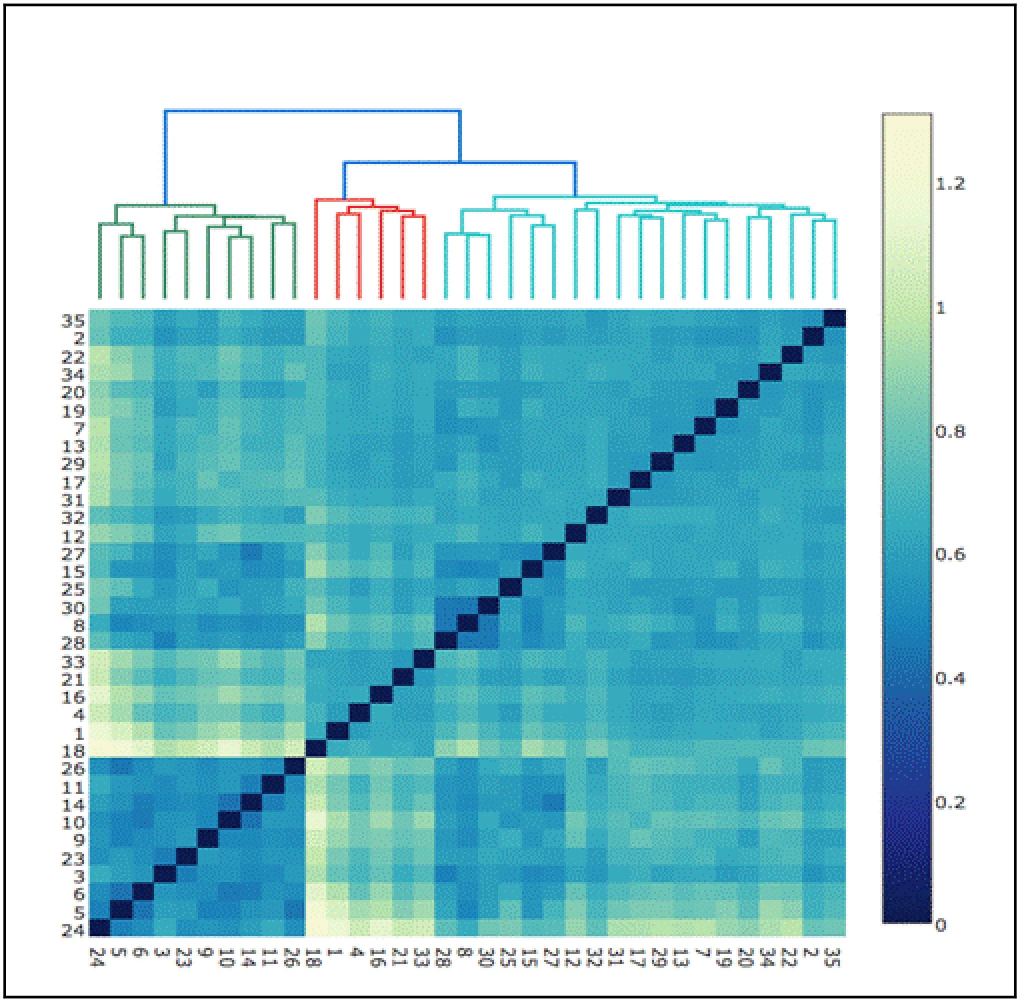
Chapter 9: Advanced Topic Modeling




Perplexity (hold_out)
Perplexity (test)
Coherence (u_mass)
Coherence (c_v)
Diff (kullback_leibler)
Convergence (jaccard)
Epochs


Fantasy
Fantastic Four (2005)[(0, 0.986)
Action
Thriller
Political Drama
Romance
Crime
Sci-Fi

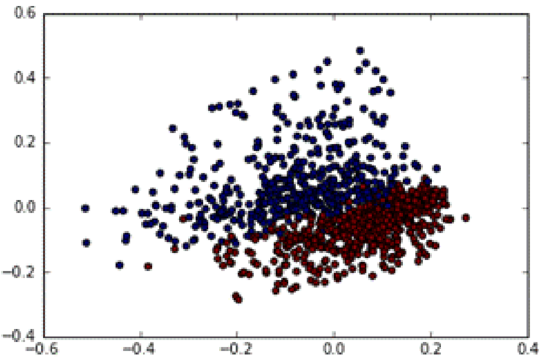

20
15
10
5
0

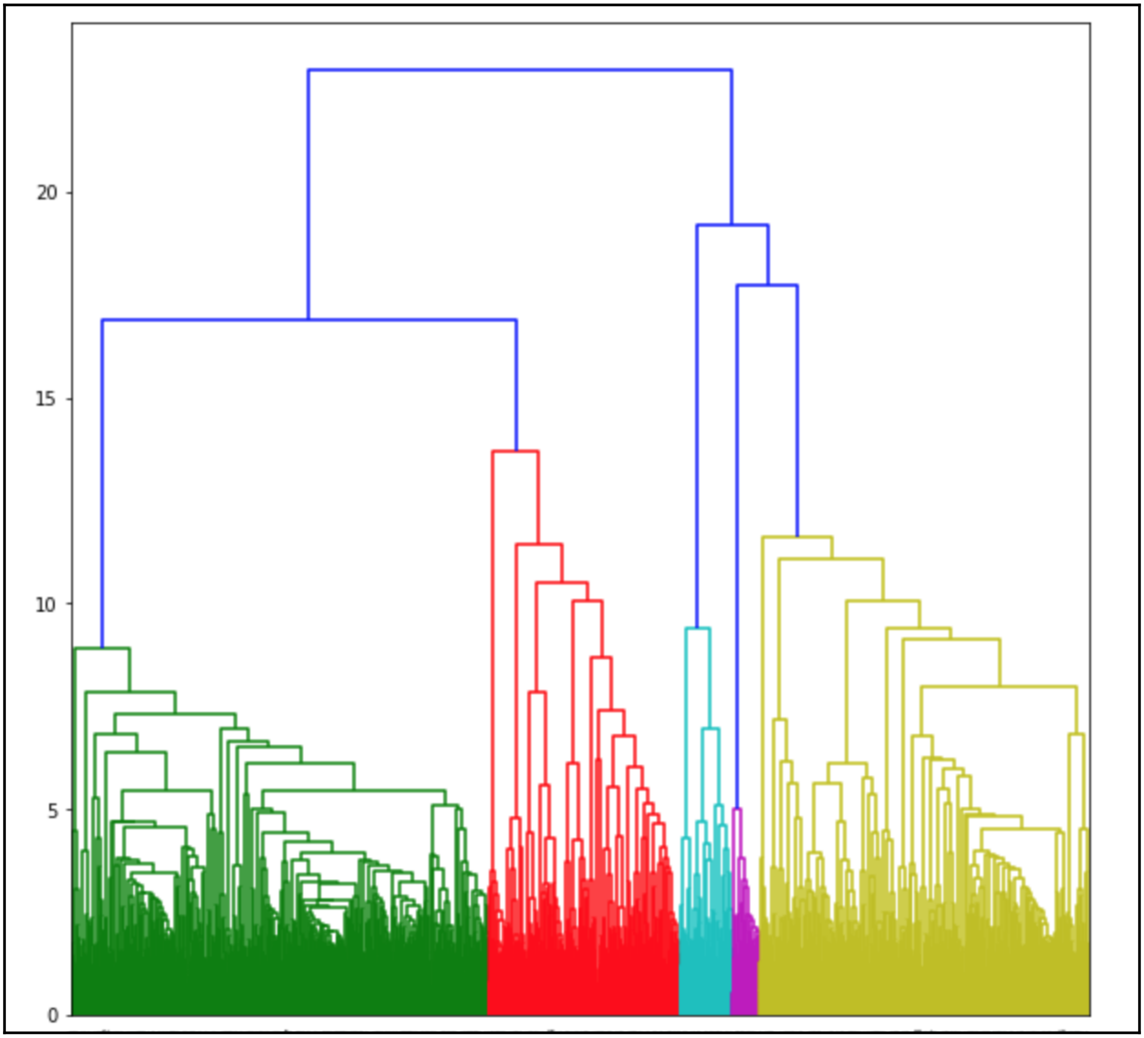

35
2
22
34
20
19
7
13
29
17
31
32
12
27
15
25
30
8
28
33
21
16
4
1
18
26
11
14
10
9
23
3
6
5
24
1.2
1
0.8
0.6
0.4
0.2
0

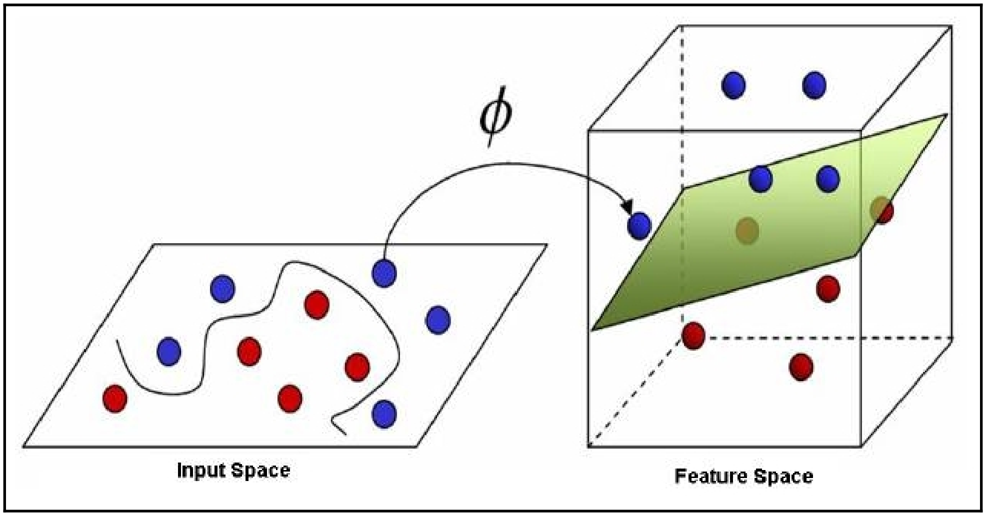
Chapter 10: Clustering and Classifying Text

Visualising Dataset

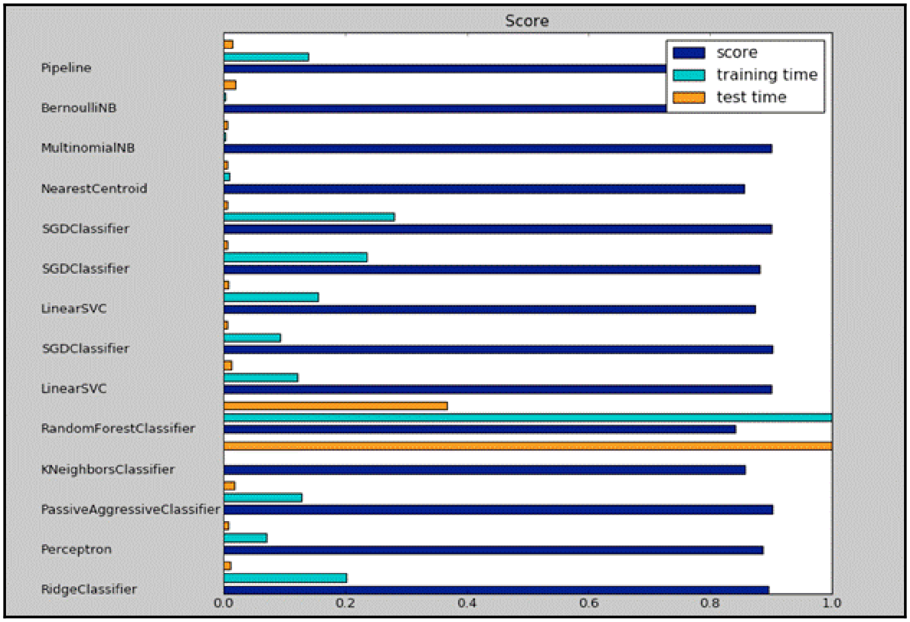
```
In [9]: from sklearn.decomposition import PCA


        newsgroups_train = fetch_20newsgroups(subset='train',
                                              categories=['alt.atheism', 'sci.space'])
        pipeline = Pipeline([
            ('vect', CountVectorizer()),
            ('tfidf', TfidfTransformer()),
        ])
        X = pipeline.fit_transform(newsgroups_train.data).todense()

        pca = PCA(n_components=2).fit(X)
        data2D = pca.transform(X)
        plt.scatter(data2D[:,0], data2D[:,1], c=newsgroups_train.target)

Out[9]: <matplotlib.collections.PathCollection at 0x10a7bae90>
```




20
15
10
5
0



ϕ
Input Space
Feature Space



Score
score
training time
test time
Pipeline
BernoulliNB
MultinomialNB
NearestCentroid
SGDClassifier
SGDClassifier
LinearSVC
SGDClassifier
LinearSVC
RandomForestClassifier
KNeighborsClassifier
PassiveAggressiveClassifier
Perceptron
RidgeClassifier
0.0
0.2
0.4
0.6
0.8
1.0

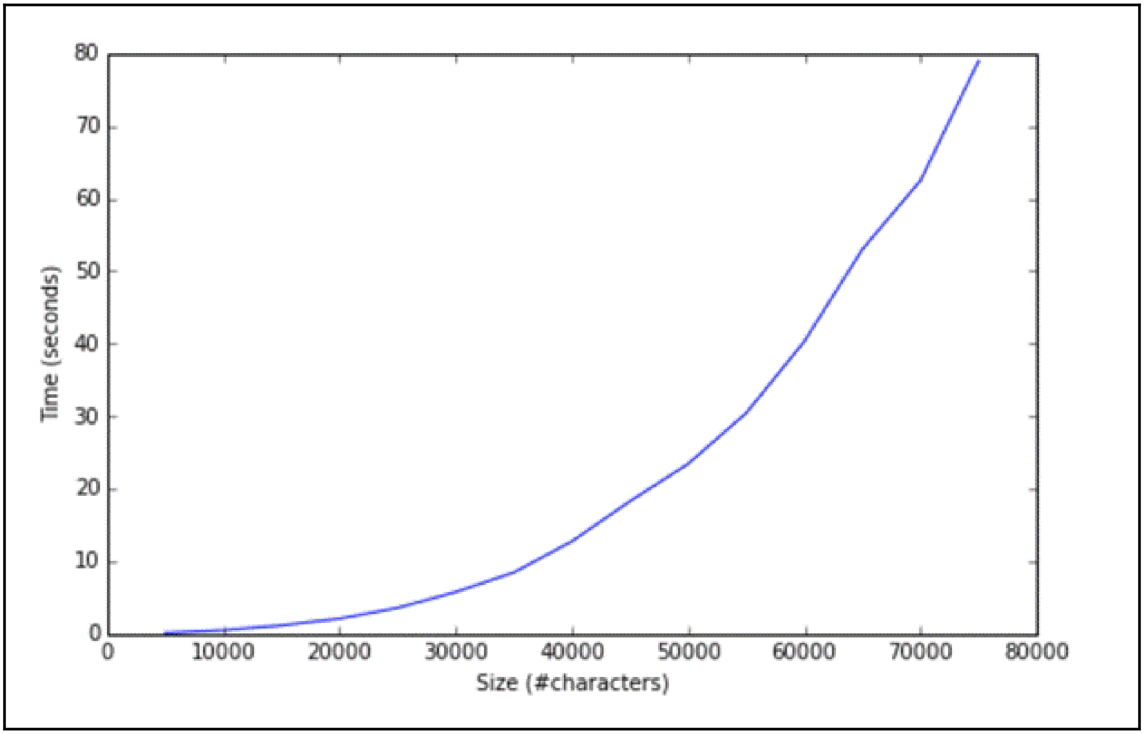
Chapter 11: Similarity Queries and Summarization

A **metric** on a set X is a function (called the *distance function* or simply **distance**)

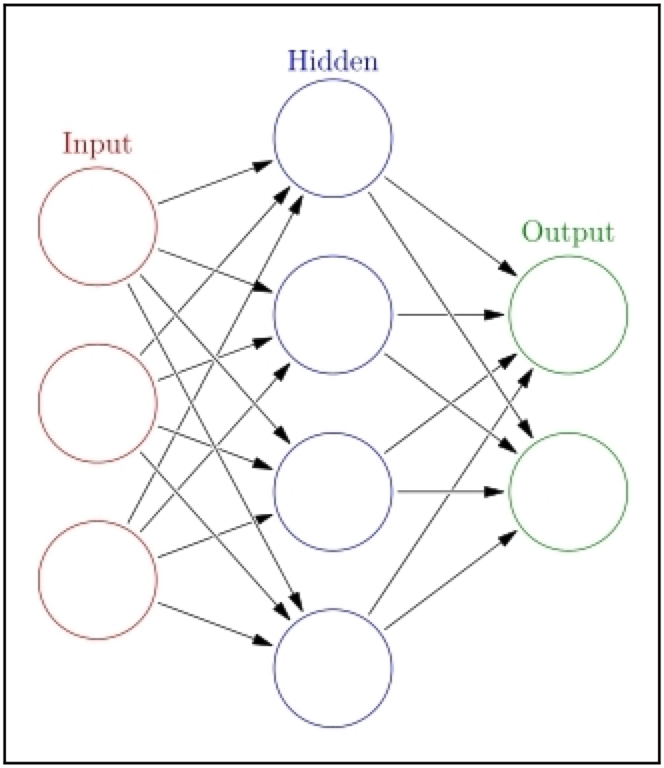
$$d : X \times X \to [0, \infty),$$

where $[0, \infty)$ is the set of non-negative real numbers and for all $x, y, z \in X$, the following conditions are satisfied:

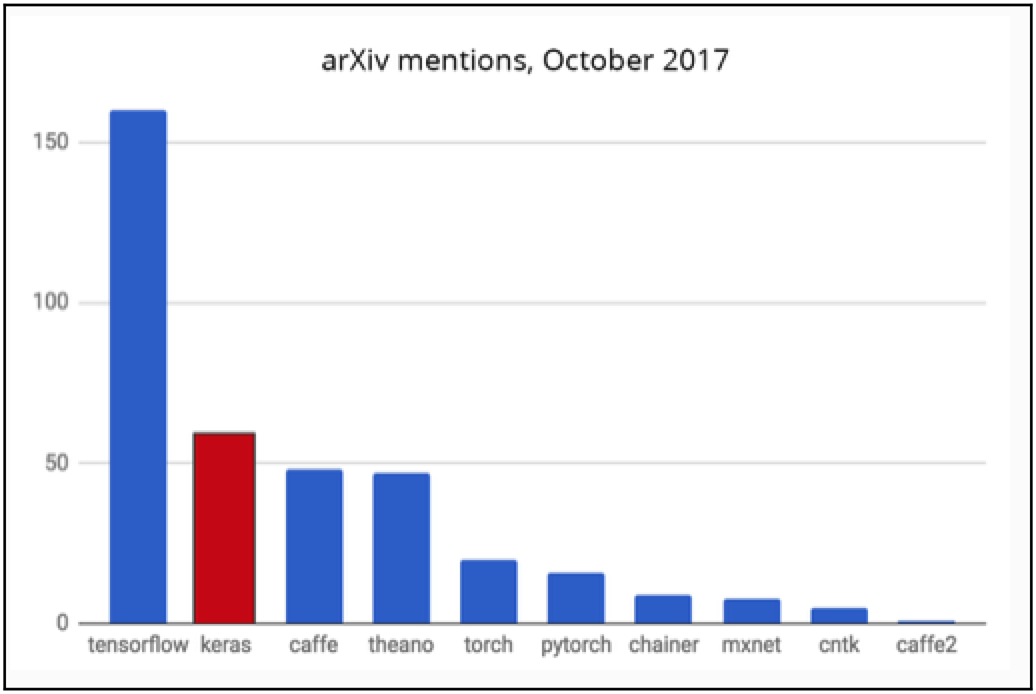
1.	$d(x, y) \geq 0$	non-negativity or separation axiom
2.	$d(x, y) = 0 \Leftrightarrow x = y$	identity of indiscernibles
3.	$d(x, y) = d(y, x)$	symmetry
4.	$d(x, z) \leq d(x, y) + d(y, z)$	subadditivity or triangle inequality



Chapter 13: Deep Learning for Text



Chapter 14: Keras and spaCy for Deep Learning



Chapter 15: Sentiment Analysis and ChatBots