I256:
Applied Natural Language Processing

Marti Hearst
Week 2

HOW DO YOU GET TO KNOW A TEXT COLLECTION?

Today's Preparatory Exercise

- Monty Python Text
  - What was unusual?
  - What did we do to it?
    - Called **Text Normalization**
    - Chris Little's code:

```
import string

text_norm =
[w.lower() for w in text6 if w[0] not in string.punctuation]

mp_norm_fd = FreqDist(text_norm)

sorted([w for w in mp_norm_fd if len(w)>5 and mp_norm_fd[w]>5])
```

LARGE DATA EXAMPLE

"principle" vs "principal"

Principle vs. Principal

prin·ci·ple
/ˈprinsəpəl/

noun

1. a fundamental truth or proposition that serves as the foundation for a system of belief or behavior or for a chain of reasoning.
   "the basic principles of Christianity"
   synonyms: truth, proposition, concept, idea, theory, assumption, fundamental, essential, ground rule
   "elementary principles"

2. a fundamental source or basis of somet
   "the first principle of all things was wate

prin·ci·pal
/ˈprinsəpəl/

adjective

1. first in order of importance; main.
   "the country's principal cities"
   synonyms: main, chief, primary, leading, foremost, first, first-line, most important, predominant, dominant, (most) prominent; More

2. (of money) denoting an original sum invested or lent.
   "the principal amount of your investment"

noun

1. the person with the highest authority or most important position in an organization, institution, or group.
   "a design consultancy whose principal is based in San Francisco"
   synonyms: chief, chief executive (officer), CEO, president, chairman, chairwoman, director, managing director, manager, head; *informal* boss
   "the principal of the firm"

2. a sum of money lent or invested on which interest is paid.
   "the winners are paid from the interest without even touching the principal"
   synonyms: capital (sum), debt, loan
   "repayment of the principal"

Large Data Example: “principle” vs “principal”

In [1]:
```
import nltk
from nltk.corpus import brown
```

First extract relevant sentences from the Brown Corpus.

In [12]:
```
principles = [' '.join(sent) for sent in brown.sents() if 'principle' in sent]
principals = [' '.join(sent) for sent in brown.sents() if 'principal' in sent]
print(str(len(principles)))
print(str(len(principals)))
```

```
104
83
```

Next, write these to a file, so we can read them back into a Text object.

In [14]:
```
of = open('palsples.txt', "w")
for sent in principles:
    of.write(sent)
for sent in principals:
    of.write(sent)
of.close
f=open('palsples.txt','rU')
raw=f.read()
tokens = nltk.word_tokenize(raw)
text = nltk.Text(tokens)
```

Now that we have a Text object, we can view words of interest as Concordances.

Now that we have a Text object, we can view words of interest as Concordances.

In [17]:
```
text.concordance('principle', 100, 50)
```

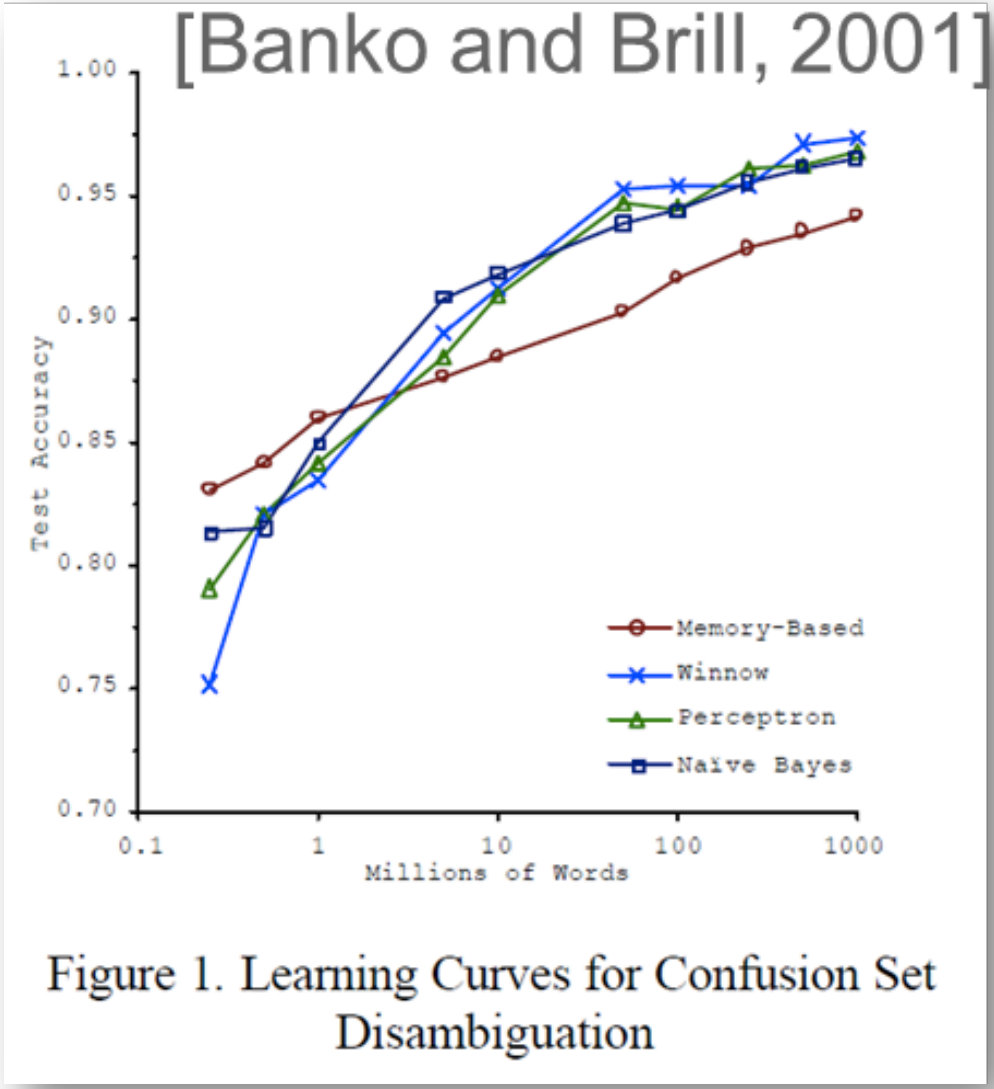
```
Displaying 50 of 107 matches:
                                                principle to the executive organs of all international
impolitic to oppose collective bargaining in principle .The Wagner Act , the Taft-Hartley Act and t
 and the Landrum-Griffin Act all endorse the principle of collective bargaining .One definition of
ing .One definition of paternalism is `` The principle or practice , on the part of a government ,
f a father dealing with his children '' .The principle is commendable but we suspect that in the pr
t of self-determination is also an essential principle .In all the bitter in-fighting , the squabbl
n from Burma showed himself both as a man of principle and a skilled diplomat .Nonetheless , althou
ch suggests that there is some kind of vital principle embodied in their faith .introduction of the
ir faith .introduction of the `` dialogue '' principle proved strikingly effective at the thirty-fo
kend in October .That Aristotelean-Thomistic principle experienced a thorough going-over from a num
e which is contained in the reciprocal trade principle under which we now operate . `` You see , fi
, the unshakeable American commitment to the principle of unconditional surrender : The tendency to
 For Wives '' was interpreted according to a principle that is becoming increasingly common in the
s n't conflict with experiments on which the principle of conservation of matter and energy is base
his own liberty rather than for any abstract principle connected with it -- such as `` cause '' .To
 , it is sufficient to point out that if the principle in terms of which alternatives are to be con
t , and if it works , accept it as a guiding principle .The Yang , or male principle , was the sour
t as a guiding principle .The Yang , or male principle , was the source of light , heat , and dynam
d with the Sun ; ; while the Yin , or female principle , flourished in darkness , cold , and quiet
ast their associations , have repudiated the principle in such clauses .It was indeed a remarkable
e of bridge building should have applied the principle of the arch , which appears in his famous br
 and Philadelphia .In following this general principle , Mason provides the observer with a natural
nstantly created .Do you try to maintain the principle of employee-contributed ( as opposed to full
fully company-paid ) programs ? ? This basic principle , the first in a richly knotted bundle , was
ound in the Nazi outlook , which contained a principle separate from and far worse than anti-Semiti
te from and far worse than anti-Semitism , a principle by which the poison of anti-Semitism itself
ent .Perhaps under the guidance of this Nazi principle one could , as Eichmann declared , feel pers
 them , the world over , operate on the same principle by which justice is administered in France a
es during adolescence involves an epigenetic principle -- during adolescence , the individual 's po
e two characteristics which violate the life principle of congregations of the major denominations
entity and must conform , therefore , to the principle of economic integration .association by leve
tion by level of achievement is the dominant principle of informal relations .The expense of this t
ned , raises fundamental questions about the principle of Protestant survival in a mobile society ;
y applications of economic pressure would in principle already have justified that use of economic
 , was coined by Bentham ) , a body of legal principle which by and large was made up of what Weste
ns could do in the world arena .That guiding principle of the Hoover Administration fell to the sie
 of the Kennedy Administration came when the principle of national responsibility for local economi
er surplus .The Rule of Law , historically a principle according everyone his `` day in court '' be
 now live in welfare states , the organizing principle of which is collective responsibility for in
ll-being .Whether a concept analogous to the principle of internal responsibility operates in a nat
- may help to reveal the contours of the new principle . ) The concept of nationalism is the politi
 The concept of nationalism is the political principle that epitomizes and glorifies the territoria
structure .Almost febrile in intensity , the principle has become worldwide in application -- unfor
k greatest harm .Complementing the political principle of nationalism is the legal principle of sov
itical principle of nationalism is the legal principle of sovereignty .The nation-state , then , ex
```

```
In [18]: text.concordance('principal', 100, 50)
Displaying 50 of 88 matches:
teacher , who defeated Felix Bush , a school principal and chairman of the Miller County Democratic
mittee .Dr . Clark has served as teacher and principal in Oklahoma high schools , as teacher and at
College .These , he said , are `` two of the principal underlying causes for family breakups leadin
s for family breakups leading to ADC '' .The principal tactic in controlling the ball was giving it
t to Abner Haynes , the flashy halfback .The principal speaker will be Senator Stuart Symington , D
e completed conversations with all the other principal Allied leaders .The principal of the school
 all the other principal Allied leaders .The principal of the school announced that -- despite the
 and mathematics .South Philadelphia High 's principal added that the current delay was caused by t
the toneless lad was making .There were four principal ones .We submit that this is a most desirabl
esirable effect of the law -- and one of its principal aims .More , the U.S. action was hailed by a
aims .More , the U.S. action was hailed by a principal opposition leader , Dr. Juan Bosch , as havi
y troubles in the near future '' .Two of the principal addresses were delivered by prominent Protes
e could see that the U.S. was in reality the principal .We had a couple of schools in this country
ad a couple of schools in this country , the principal one being on the Marshall Field estate out i
rmer Governor General of Australia , was the principal British commander in the field during the Bu
 Burma War .Last season , the Comedie 's two principal experiments came to grief , and , in consequ
witches back to a consideration of the seven principal Presidential hopefuls : five Democrats -- Se
on the precision and incisiveness of the two principal combatants .However , my principal objection
f the two principal combatants .However , my principal objection in this sort of novel is to the ha
 population shifts have emptied churches , a principal reason for this phenomenon of redundancy is
ns of the Han period and the latter from the principal river systems of Old China .But in such an i
d be satisfied if the judgment were that the principal objection to the identity of forces which pr
ing which is contrary to it .There are three principal feed bunk types for dairy and beef cattle :
as the Tri-State Pipeline Corporation , with principal offices in New York State .Tri-State has acq
ts exclusive distribution for the northern , principal heating states .If it is owned , taxes must
, there will be interest and payments on the principal to take care of .The role of an earthquake i
y frightening , but fire may actually be the principal agent in a particular disaster .Oils , or li
he seeds of flax and tung have long been the principal constituents of paints and varnishes for pro
ntries where cereal grains are not among the principal crops of a region , starchy tubers or roots
nomic integration through co-optation is the principal form of mission in the contemporary church ;
 furnish presents to hold the loyalty of the principal Indians .For southeastern Louisiana , Mobile
.For southeastern Louisiana , Mobile was the principal post , and it was to furnish supplies for tr
British traders .On the middle Mississippi a principal post was to be located near the mouth of the
 the mouth of the Arkansas .Each of the five principal posts was to have a director , responsible t
 a director-general at New Orleans .Only two principal storehouses were actually established -- one
ther at New Orleans .The Chickasaws were the principal source of trouble in the Mobile district .Th
ource of trouble in the Mobile district .The principal mauler , however , was Senator Joseph McCart
 poplar trees that now serves as the city 's principal street .The principal defender of this view
 serves as the city 's principal street .The principal defender of this view of primary experience
ficacy '' is Alfred North Whitehead .but his principal theme is that the intrigues of the Tories ,
iate threat to Church and State .This is the principal point made in this final section of Englishm
he body of the clergy .In the summary of the principal events of the campaign compiled from the off
of the Cossack Corps , had invested Islam 's principal stronghold on the north shore of the Black S
o complicated that even Nogaret , one of the principal actors in the drama , could misinterpret the
ies to a more immediate cause .Of course the principal factor in the whole experience was the kind
tion he received .SBA works closely with the principal property disposal installations of the Feder
```

“principle” vs “principal”

- “principle of conservation”

- “served as a teacher and a principal”



Lessons:

More data does better than smarter algorithms
AND
Results keep improving with more data.

HOW DO YOU GET TO KNOW A TEXT COLLECTION?

Example: Wolfram Alpha

- Consolidates data from many databases
- What happens when you type in the name of an open source (public domain) collection?

WolframAlpha computational... knowledge engine

tale of two cities

Examples Random

Assuming "tale of two cities" is a novel | Use as a movie instead

Input interpretation:

A Tale of Two Cities

Basic novel information:

full title	A Tale of Two Cities
document type	novel
author	Charles Dickens
first publication date	April 30, 1859 (155 years ago)
publisher	All The Year Round (serial) | Chapman & Hall (book)
original language	English

Image:

Opening phrase:

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way––in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

Word properties: Show distribution

number of words	136 955 words (silent reading: 8 hours)
number of unique words	9690 words
number of unique word stems	6569
average word length	4.25 characters
longest word	undistinguishable

Most frequent words: More

the (5.84%) | and (3.65%) | of (2.92%) | to (2.6%) | a (2.14%) | in (1.89%) | it (1.51%) | his (1.47%) | that (1.42%) | I (1.42%) | he (1.35%) | was (1.29%) | you (1.02%) | with (0.96%) | had (0.95%) | as (0.85%) | her (0.76%) | at (0.75%) | him (0.71%) | for (0.7%) | ...

Most frequent capitalized words: More

Mr (0.45%) | Lorry (0.27%) | Defarge (0.22%) | Manette (0.12%) | Pross (0.12%) | Carton (0.11%) | Darnay (0.11%) | Lucie (0.09%) | Cruncher (0.09%) | Stryver (0.08%) | Jerry (0.08%) | Monseigneur (0.08%) | Charles (0.07%) | Tellson (0.07%) | Sydney (0.06%) | Jacques (0.05%) | Paris (0.05%) | France (0.04%) | Saint (0.04%) | Antoine (0.04%) | ...

Most frequent two-word phrases: More

of the (922 times) | in the (715 times) | to the (407 times) | and the (335 times) | Mr Lorry (325 times) | on the (321 times) | it was (318 times) | at the (292 times) | to be (286 times) | he had (279 times) | in his (245 times) | in a (245 times) | it is (234 times) | with a (223 times) | of his (217 times) | ...

Sentence properties: Show longest Show distribution

number of sentences	7580 sentences
average sentence length	97.93 characters
	18.07 words

Paragraph properties: Show distribution

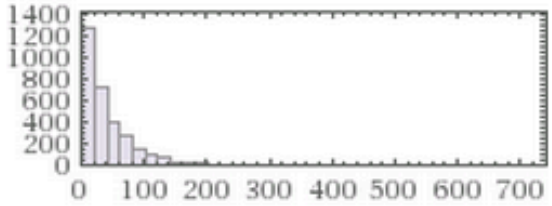
number of paragraphs	3257 paragraphs
average paragraph length	231.1 characters
	42.03 words
	2.32 sentences

Paragraph properties:

Hide distribution

number of paragraphs	3257 paragraphs
average paragraph length	231.1 characters
	42.03 words
	2.32 sentences

Words in paragraph distribution:



Readability:

automated readability index	7.6 (estimated US grade level)
Coleman–Liau index	7.6 (estimated US grade level)

Definitions »

Wolfram Alpha Book Stats

- This is one way to give an overview of a text
- Ch. 1 of NLTK book supplies much of what's needed