# Speech and Text Processing

- Dan Jurafsky
- James Martin

# Chapter 2 : Regular Expressions , Text Normalization , Edit Distance

- Text Normalization
  - → tokenizing
  - → lemmatization
  - → Stemming
  - → Sentence segmentation

# Exercises

(Q 2.1)

(a) `\b[a-zA-Z]+\b`

(b) `\b[a-z]*b\b`

(c) `\b b+(ab+)+\b`

(Q 2.2)

(a) `(.+)␣\1`

(b) `^[0-9]+.*[A-Za-z]*\.`

(c) `\bgrotto\b.*\braven\b|`
`\braven\b.*\bgrotto\b`

(Q 2.4)

| | # | d | e | a | l |
|---|---|---|---|---|---|
| # | 0 | 1 | 2 | 3 | 4 |
| l | 1 | 1 | 2 | 3 | 4 |
| e | 2 | 2 | 1 | 2 | 3 |
| d | 3 | 2 | 2 | 2 | 3 |
| a | 4 | 3 | 3 | 3 | 3 |

(Q 2.5)

| | # | b | r | i | e | f |
|---|---|---|---|---|---|---|
| # | 0 | 1 | 2 | 3 | 4 | 5 |
| d | 1 | 2 | 3 | 4 | 5 | 6 |
| r | 2 | 3 | 2 | 3 | 4 | 5 |
| i | 3 | 4 | 3 | 2 | 3 | 4 |
| v | 4 | 5 | 4 | 3 | 4 | 5 |
| e | 5 | 6 | 5 | 4 | 5 | 6 |

| | # | d | i | v | e | r | s |
|---|---|---|---|---|---|---|---|
| # | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| d | 1 | 0 | 1 | 2 | 3 | 4 | 5 |
| r | 2 | 1 | 2 | 3 | 4 | 5 | 6 |
| i | 3 | 2 | 1 | 2 | 3 | 4 | 5 |
| v | 4 | 3 | 2 | 1 | 2 | 3 | 4 |
| e | 5 | 4 | 3 | 2 | 1 | 2 | 3 |

# Chapter 3 : N-gram Language Models

## Exercises

(Q 3.1) $P(w_n | w_{n-2} w_{n-1}) = \dfrac{C(w_{n-2} w_{n-1} w_n)}{C(w_{n-2} w_{n-1})}$

<s><s> I am Sam </s>
<s><s> Sam I am </s>
<s><s> I do not like green eggs and ham </s>

$P(I \mid \langle s\rangle\langle s\rangle) = 2/3$
$P(am \mid \langle s\rangle I) = 1/2$

(Q 3.2) $P(\text{i want chinese food}) = P(i|\langle s\rangle)\, P(want|i)\, P(chinese|want)\, P(food|chinese)\, P(\langle /s\rangle|food)$

$= 0.25 \times .33 \times 0\ 0065 \times .52 \times 0.68$
$= 0.0001896$

$P(\text{i want chinese food}) = P(i|\langle s\rangle)\, P(want|i)\, P(chinese|want)\, P(food|chinese)\, P(\langle /s\rangle|food)$

$= .19 \times 0.21 \times 0.0029 \times 0.052 \times .4$
$= 0.000002406$

(Q 3.3) The unsmoothed probability is higher because the bigrams used in the sentences are very common and has probablities. However, in the smoothed case, their probablities are distributed among not-so-common bigrams which are not used in our test statement.

(Q 3.4)

| | <s> | I | am | Sam | do | not | like | green | eggs | and | </s> |
|---|---|---|---|---|---|---|---|---|---|---|---|
| <s> | 1 | 4 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| I | 1 | 1 | 4 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 |
| am | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| sam | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 4 |
| do | " | 1 | " | " | " | 2 | " | " | " | " | 1 |
| not | " | " | " | " | " | 1 | 2 | " | " | " | " |
| like | " | " | " | " | " | " | 1 | 2 | " | " | " |
| green | " | " | " | " | " | " | " | 1 | 2 | " | " |
| eggs | " | " | " | " | " | " | " | " | 1 | 2 | " |
| and | " | " | " | 2 | " | " | " | " | " | 1 | " |
| </s> | " | " | " | 1 | " | " | " | " | " | " | " |

$$P(\text{Sam} \mid \text{am}) = \frac{C(\text{am Sam})}{C(\text{am})} = \frac{3}{14} = .214$$

(Q 3.5)

| | <s> | a | b |
|---|---|---|---|
| <s> | 0 | 2 | 2 |
| a | 0 | 1 | 1 |
| b | 0 | 1 | 1 |

- $P(a\,a) = P(a \mid \langle s\rangle)\,P(a \mid a) = 0.5 \times 0.5 = 0.25$
  $P(a\,b) = P(a \mid \langle s\rangle)\,P(b \mid a) = 0.5 \times 0.5 = 0.25$
  $P(b\,b) = P(b \mid \langle s\rangle)\,P(b \mid b) = 0.5 \times 0.5 = 0.25$
  $P(b\,a) = P(b \mid \langle s\rangle)\,P(a \mid b) = 0.5 \times 0.5 = 0.25$

$P(s \in \{a,b\}^2) = 1$

(Q3.6) $P(w_3 | w_1 w_2) = \dfrac{C(w_1 w_2 w_3) + 1}{\sum_{w} (C(w_1 w_2 w) + 1)}$

$= \dfrac{C(w_1 w_2 w_3) + 1}{C(w_1 w_2) + V}$

(Q3.7) = $P(\text{sam} | \text{am}) = \lambda_1 P(\text{Sam}) + \lambda_2 (\text{Sam} | \text{am})$

$= \dfrac{1}{\cancel{2}} \times \dfrac{4\cancel{2}}{25} + \dfrac{1}{\cancel{2}} \times \dfrac{\cancel{2}}{3}$

$= \dfrac{2}{25} + \dfrac{1}{3} = 0.41$

(Q3.12) $PP(W) = \sqrt[N]{\prod_{i=1}^{N} \dfrac{1}{P(w_i)}}$

$P(0) = \dfrac{91}{100}$

$P(1) = P(2) \ldots P(9) = \dfrac{1}{100}$

$= \sqrt[10]{\dfrac{(100)^{10}}{(91)^{9}}}$

$= 1.726$

# Chapter 4: Naive Bayes and Sentiment Classification

(Q4.1) $S$ = "I always like foreign films"

$$P(neg|S) = \frac{P(S|neg)\,P(neg)}{P(S)}$$

$$P(pos|S) = \frac{P(S|pos)\,P(pos)}{P(S)}$$

ignore common base.

- $p(neg|S) = (0.16 \times 0.06^2 \times 0.15 \times 0.11) \times 0.5$

  $= 0.000004752$

- $P(pos|S) = (0.09 \times 0.07 \times 0.29 \times 0.04 \times 0.08) \times 0.5$

  $= 0.000002923$

The naive bayes will assign "neg" class to the sentence because $P(neg|S) > P(pos|S)$

(Q4.2) $P(\text{comedy}) = 2/5$ $\quad |V| = 7$

$P(\text{action}) = 3/5$

$$P(\text{fast} \mid \text{comedy}) = \frac{\text{count}(\text{fast}, \text{comedy}) + 1}{\sum_{w \in V} (\text{count}(w, \text{comedy}) + 1)}$$

$$= \frac{2}{9+7} = \frac{2}{16}$$

$$P(\text{fast} \mid \text{action}) = \frac{3}{11+7} = \frac{3}{18}$$

$P(\text{couple} \mid \text{comedy}) = \frac{3}{16}$ $\quad P(\text{shoot} \mid \text{comedy}) = \frac{1}{16}$

$P(\text{couple} \mid \text{action}) = \frac{1}{18}$ $\quad P(\text{shoot} \mid \text{action}) = \frac{5}{18}$

$P(\text{fly} \mid \text{comedy}) = \frac{2}{16}$

$P(\text{fly} \mid \text{action}) = \frac{2}{18}$

$$P(\text{comedy} \mid D) = P(D \mid \text{comedy})\, P(\text{comedy})$$
$$= \frac{2}{16} \times \frac{3}{16} \times \frac{2}{16} \times \frac{1}{16} \times \frac{2}{5}$$

$$= 0.0000732$$

$$P(\text{action} \mid D) = 3/18 \times 1/18 \times 2/18 \times 5/18 \times 3/5$$
$$= 0.000171$$

D will be classified as "action".

(Q 4.3) • Binarized naive Bayes

$P(neg) = 0.6$ $\quad$ $P(pos) = 0.4$

$P(good|neg) = 3/9$ $\quad$ $P(good|pos) = 2/7$

$P(poor|neg) = 4/9$ $\quad$ $P(poor|pos) = 2/7$

$P(great|neg) = 2/9$ $\quad$ $P(great|pos) = 3/7$

$$P(neg|D) = \frac{3}{9} \times \frac{4}{9} \times \frac{2}{9} \times 0.6 = 0.0197$$

$$P(pos|D) = \frac{2}{7} \times \frac{2}{7} \times \frac{3}{7} \times 0.4 = 0.0139$$

Classified as "neg" by BNB.

• Multinomial Naive Bayes

$P(good|pos) = 4/12$ $\quad$ $P(good|neg) = 3/17$

$P(poor|pos) = 2/12$ $\quad$ $P(poor|neg) = 11/17$

$P(great|pos) = 6/12$ $\quad$ $P(great|neg) = 3/17$

$$P(pos|D) = \left(\frac{4}{12}\right)^2 \times \cancel{\frac{2}{12} \times \frac{6}{12}} \times 0.6$$

$$= 0.0055$$

$$P(neg|D) = \left(\frac{3}{17}\right)^2 \times \left(\frac{11}{17}\right) \times \left(\frac{3}{17}\right) \times 0.4$$

$$= 0.0014$$

Classified as "pos" by MNB.
Both models disagree.

# Chapter 10:

CQ 10.1) REF: witnessthe past (18 unigram, 17 bigrams)
HYP2: pastwitness (11 unig., 10 bigrams)

Unigrams that match: pastwitness (11 unigrams)
bigrams that match: pa as st wi it tn ne es ss (9 bigrams)

Unigram P = 11/11 Unigram R: 11/18
Bigram P = 9/10 Bigram R: 9/17

$$\text{chrP} = (11/11 + 9/10)/2 = .95$$

$$\text{chrR} = (11/18 + 9/17)/2 = 0.57$$

$$\text{Chrf}_{2,2} = 5\frac{\text{chrP} * \text{chrR}}{4\,\text{chrP} + \text{chrR}} = \frac{5 \times .95 \times 0.57}{4 \times .95 + 0.57}$$

$$= 0.6199$$

# Chapter 12: Constituency Grammars

(Q 12.1)

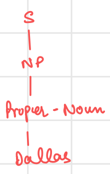
Grammar used:

S → NP | VP | NP VP | NP PP
NP → Pronoun | Proper-Noun | Det Nominal | Nominal | Adj
Nominal → Nominal N | Adj N | N
VP → V | V NP | V NP PP | V PP
PP → Preposition NP

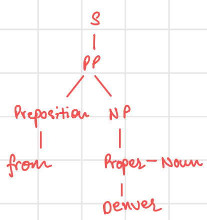
Lexicals used:

Adj → early | all | one-way | any
N → p.m | flights | redeye | fare | delays | five
Det → a
V → arriving
Preposition → from | in | after | on
Proper-Noun → Denver | Dallas | Washington | Thrusday

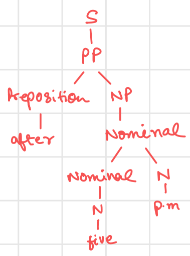
(a) Dallas



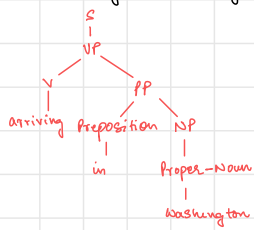

(b) from Denver



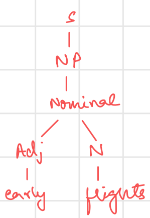

(c) after five p.m



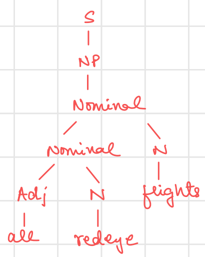

(d) arriving in washington



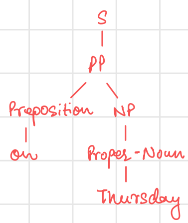

(e) early flights



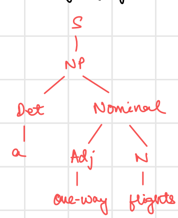

(f) all redeye flights



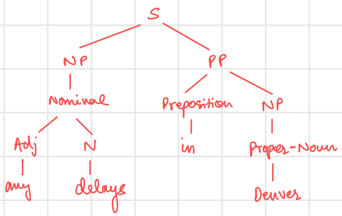

(g) on Thursday



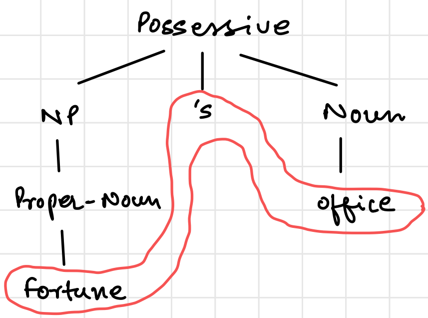

(h) a one-way -flight

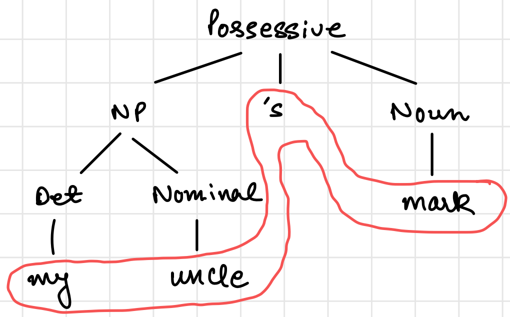


(i) any delays in denver




(Q12.6) Possessive $\rightarrow$ NP 's Noun




(Q12.8)

- Step 1: Remove all terminals from right hand side if they occur with a non-terminal.

Example:

$A \rightarrow bC \implies A \rightarrow BC$
$B \rightarrow b$

- Step 2: on rules with more than 2 non-terminals on RHS, create new rules and substitute.

Example: $A \rightarrow BCD \implies A \rightarrow BE$
$E \rightarrow CD$