Lecture 3-1 overview

• Today: **Feature extraction** from text.
  • How to pick the right features?
  • Grammatical ‘parts-of-speech’.
    • (which don’t require spoken language)

• Some slides *may* be based on content from Bob Carpenter, Dan Klein, Roger Levy, Josh Goodman, Dan Jurafsky, and Christopher Manning.
Features

• **Feature**: *n.* A measurable **variable** that is (rather, *should be*) **distinctive** of something we want to model.

• We usually choose features that are useful to **identify** something, i.e., to do **classification**.
  - E.g., the **number of profanities** in a Tweet was once useful in identifying whether a Twitterer is a President or not.

• We often need **several** features to adequately model something – *but not too many!*
Feature vectors

• Values for several features of an observation can be put into a single vector.

<table>
<thead>
<tr>
<th># proper nouns</th>
<th># 1st person pronouns</th>
<th># commas</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Feature vectors

- Features should be useful in discriminating between categories.

Table 3: Features to be computed for each text

- Counts:
  - First person pronouns
  - Second person pronouns
  - Third person pronouns
  - Coordinating conjunctions
  - Past-tense verbs
  - Future-tense verbs
  - Commas
  - Colons and semi-colons
  - Dashes
  - Parentheses
  - Ellipses
  - Common nouns
  - Proper nouns
  - Adverbs
  - "wh-words"
  - Modern slang acronyms
  - Words all in upper case (at least 2 letters long)
- Average length of sentences (in tokens)
- Average length of tokens, excluding punctuation tokens (in characters)
- Number of sentences

**Higher values** → this person is referring to themselves (to their opinion, too?)

**Higher values** → looking forward to (or dreading) some future event?

**Lower values** → this tweet is more formal. Perhaps not overly sentimental?
Quick comment on noise

• **Noise** is generally any **artifact** in your received ‘**signal**’ that **obfuscates** (hides) the features you want.

• E.g., in **acoustics**, it can be a loud buzzing sound that washes out someone’s voice.

• E.g., in **tweets**, it can be text that invalidates your counts.
  • E.g., The semi-colon in “… octopus ;)” is part of an **emoticon**; will it confuse our **classifier** if we count it as punctuation?

**Note**: you don’t have to deal with emoticons in A1.
Pre-processing

• **Pre-processing** involves preparing your data to make feature extraction easier or more valid.
  • E.g., punctuation likes to press up against words. The sequence “example,” should be counted as **two** tokens – not one.
  • We separate the punctuation, as in “example,”.

• **There is no perfect pre-processor.**
  Mutually exclusive approaches can often **both** be justified.
  • E.g., Is Newfoundland-Labrador **one** word type or **two**?
    Each answer has a unique implication for splitting the dash.
  • Often, **noise-reduction** removes **some** information.
  • Being **consistent** is important.
Determining a good set of features

• Restricting your feature set to a proper subset quickens **training** and reduces **overfitting**.

• There are a few methods that select good features, e.g.,
  • Correlation-based feature selection
  • Minimum Redundancy, Maximum Relevance
Pearson’s correlation

• **Correlation** is a measure of **linear** dependence

\[ \rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \]

• Does not measure ‘slope’ nor **non-linear** relations.
Correlation-based feature selection

- ‘Good’ features should correlate strongly (+ or -) with the predicted variable but not with other features.

- $S_{CFS}$ is some set $S$ of $k$ features $f_i$ that maximizes this ratio, given class $c$:

$$S_{CFS} = \arg\max_S \frac{\sum_{f_i \in S} \rho_{cf_i}}{\sqrt{k + 2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \rho_{f_if_j}}}$$
mRMR feature selection

• Minimum-redundancy-maximum-relevance (mRMR) can use correlation, distance scores (e.g., $D_{KL}$) or mutual information to select features.

• For feature set $S$ of features $f_i$, and class $c$,

$$D(S, c)$$ : a measure of relevance $S$ has for $c$, and

$$R(S)$$ : a measure of the redundancy within $S$,

$$S_{mRMR} = \arg\max_S \left[D(S, c) - R(S)\right]$$
mRMR feature selection

- Measures of **relevance** and **redundancy** can make use of our familiar measures of **mutual information**,

\[
D(S, c) = \frac{1}{||S||} \sum_{f_i \in S} I(f_i; c)
\]

\[
R(S) = \frac{1}{||S||^2} \sum_{f_i \in S} \sum_{f_j \in S} I(f_i; f_j)
\]

- mRMR is **robust** but doesn’t measure **interactions** of features in estimating \( C \) (for that we could use ANOVAs).
Different features for different tasks

- **Alzheimer’s disease** involves atrophy in the brain.
  - Excessive **pauses** (acoustic disfluencies),
  - Excessive **word type repetition**, and
  - Simplistic or **short** sentences.
    - ‘**function words**’ like the and an are often **dropped**.

- To **diagnose** Alzheimer’s disease, one might measure:
  - **Proportion** of utterance spent in silence.
  - **Entropy** of word type usage.
  - **Number** of word **tokens** in a sentence.
    - **Number of prepositions** and **determiners** (explained shortly).
Features in Sentiment Analysis

• **Sentiment analysis** can involve detecting:
  • Stress or frustration in a conversation.
  • Interest, confusion, or preferences. Useful to marketers.
    • e.g., ‘omg i got a lowzy mbp 2016 4xmas.fml’
    • Lies. e.g., ‘Let’s watch Netflix and chill.’

• Complicating factors include sarcasm, implicitness, and a subtle spectrum from negative to positive opinions.

• **Useful features** for sentiment analyzers include:
  • Trigrams.
  • First-person pronouns.
Parts of speech (PoS)

• Linguists like to group words according to their **structural function** in building sentences.
  • This is similar to grouping Lego by their shapes.

• **Part-of-speech**: *n.* lexical category or morphological class.

Nouns collectively constitute a part of speech (called **Noun**)
## Example parts of speech

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>is usually a <strong>person, place, event, or entity.</strong></td>
<td>chair, pacing, monkey, Gertrude.</td>
</tr>
<tr>
<td>Verb</td>
<td>is usually an <strong>action or predicate.</strong></td>
<td>punch, debate, explicate.</td>
</tr>
<tr>
<td>Adjective</td>
<td>modifies a <strong>noun</strong> to further describe it.</td>
<td>orange, obscene, disgusting.</td>
</tr>
<tr>
<td>Adverb</td>
<td>modifies a <strong>verb</strong> to further describe it.</td>
<td>lovingly, horrifyingly, often</td>
</tr>
</tbody>
</table>
# Example parts of speech

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preposition</td>
<td>Often specifies aspects of space, time, or means.</td>
<td>around, over, under, after, before, with</td>
</tr>
<tr>
<td>Pronoun</td>
<td></td>
<td>Look it up yourself!</td>
</tr>
<tr>
<td>Determiner</td>
<td>logically quantify words, usually nouns.</td>
<td>the, an, both, either</td>
</tr>
<tr>
<td>Conjunction</td>
<td>combines words or phrases.</td>
<td>and, or, although</td>
</tr>
</tbody>
</table>
Other parts of speech

- **Particles:** up, down, on, off
  - e.g., throw her coat off
  $$\equiv$$ throw off her coat

- **Auxiliaries:** can, may, should, is, have

- **Numerals:** one, $19.99, 6.02 \times 10^{23}$

- **Punctuation:** ), (, :, ,, .

- **Symbols:** +, %, &

- **Interjection:** uh, hmmm, duh, herp

- ...
Contentful parts-of-speech

• Some PoS convey more **meaning**.
  • Usually nouns, verbs, adjectives, adverbs.
  • **Contentful** PoS usually contain more words.
    • e.g., there are more **nouns** than **prepositions**.

• **New** contentful words are continually **added**
  e.g., an **app**, **to google**, **to misunderstand**.

• **Archaic** contentful words go **extinct**.
  e.g., **fumificate**, v., (1721-1792),
  **frenigerent**, adj., (1656-1681),
  **melanochalcogapher**, n., (c. 1697).
Functional parts-of-speech

• Some PoS are ‘glue’ that holds others together.
  • E.g., prepositions, determiners, conjunctions.
  • **Functional** PoS usually cover a **small** and **fixed** number of word types (i.e., a ‘**closed class**’).

• Their **semantics** depend on the contentful words with which they’re used.
  • E.g., *I’m on time* vs. *I’m on a boat*
Grammatical features

• There are several grammatical features that can be associated with words:
  • Case
  • Person
  • Number
  • Gender

• These features can restrict other words in a sentence.
Grammatical features – case

• **Case**: *n.* the **grammatical** form of a **noun** or **pronoun**.

• E.g.,

  - **nominative**: the **subject** of a verb (e.g., “**We** remember”)
  - **accusative**: the **direct object** of a verb
    (e.g., “**You** remember us”)
  - **dative**: the **indirect object** of a verb
    (e.g. “I gave your **mom** the book”)
  - **genitive**: indicates **possession**
    (e.g., “your **mom**’s book”)

...
Grammatical features – person

- **Person**: *n.* typically refers to a participant in an event, especially with **pronouns** in a conversation.

- E.g.,
  - **first**: The speaker/author. Can be either inclusive ("we") or exclusive of hearer/reader ("I").
  - **second**: The hearer/reader, exclusive of speaker ("you")
  - **third**: Everyone else ("they")
Grammatical features – number

- **Number**: n. Broad numerical distinction.

- E.g.,
  - **singular**: Exactly one (“one cow”)
  - **plural**: More than one (“two cows”)
  - **dual**: Exactly two (e.g., - ان in Arabic).
  - **paucal**: Not too many (e.g., in Hopi).
  - **collective**: Countable (e.g., Welsh “moch” for ‘pigs’ as opposed to “mochyn” for vast ‘pigginess’).

...
Grammatical features – gender

• **gender**: *n.* typically partitions **nouns** into classes associated with biological gender. **Not** typical in English.
  • Gender alters neighbouring words **regardless** of speaker/hearer.

• E.g.,
  - **feminine**: Typically **pleasant** things (not always).
    (e.g., *la France, eine Brücke, une poubelle*).
  - **masculine**: Typically **ugly** or **rugged** things (not always).
    (e.g., *le Quebec, un pont*).
  - **neuter**: Everything else.

*(Brücke: German bridge; pont: French bridge; poubelle: French garbage)*
Other features of nouns

- **Proper noun:** named things (e.g., “they’ve killed Bill!”)
- **Common noun:** unnamed things (e.g., “they’ve killed the bill!”)
- **Mass noun:** divisible and uncountable. (e.g., “butter” split in two gives two piles of butter – not two ‘butters’)
- **Count noun:** indivisible and countable. (e.g., a “pig” split in two does not give two pigs)
Some features of prepositions

• By
  • Alongside: a cottage by the lake
  • Agentive: Chlamydia was given to Mary by John

• For
  • Benefactive: I have a message for your mom
  • Purpose: have a friend (over) for dinner

• With
  • Sociative: watch a film with a friend
  • Instrumental: hit a nail with a hammer
Agreement

• Parts-of-speech **should** match (i.e., **agree**) in certain ways.

• **Articles** ‘have’ to **agree** with the **number** of their **noun**
  • e.g., “**these** pretzels are making me thirsty” 😊
  • e.g., “**a** winters are coming” 😊

• **Verbs** ‘have’ to **agree** (at least) with their **subject** (in English)
  • e.g., “the **dogs** eats the gravy” 😞 no **number** agreement

  • e.g., “**Yesterday**, all my troubles **seem** so far away” 😞 **bad tense** – should be past tense **seemed**
PoS tagging

• Tagging: *e.g.* the process of assigning a part-of-speech to each word in a sequence.

• *E.g.*, using the ‘Penn treebank’ tag set *(see appendix)*:

<table>
<thead>
<tr>
<th>Word</th>
<th>The</th>
<th>nurse</th>
<th>put</th>
<th>the</th>
<th>angry</th>
<th>koala</th>
<th>to</th>
<th>sleep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag</td>
<td>DT</td>
<td>NN</td>
<td>VBD</td>
<td>DT</td>
<td>JJ</td>
<td>NN</td>
<td>IN</td>
<td>NN</td>
</tr>
</tbody>
</table>
Ambiguities in parts-of-speech

• Words can belong to many parts-of-speech.
  • E.g., back:
    • The back/JJ door (adjective)
    • On its back/NN (noun)
    • Win the voters back/RB (adverb)
    • Promise to back/VB you in a fight (verb)

• We want to decide the appropriate tag given a particular sequence of tokens.
Why is tagging useful?

- First step towards practical purposes.
  - E.g.,
    - **Speech synthesis**: how to pronounce text
      - I’m *content*/JJ vs. *the content*/NN
      - I *object*/VBP vs. *the object*/NN
      - I *lead*/VBP (“l iy d”) vs. *it’s lead*/NN (“l eh d”)
    - **Information extraction**: Quickly finding names and relations.
    - **Machine translation**: Identifying grammatical ‘chunks’ is useful.
Tagging as classification

- We have access to a **sequence of observations** and are expected to decide on the best assignment of a **hidden variable**, i.e., the PoS

<table>
<thead>
<tr>
<th>Hidden variable</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PRP</strong></td>
<td><strong>she</strong></td>
</tr>
<tr>
<td><strong>VBD</strong></td>
<td><strong>promised</strong></td>
</tr>
<tr>
<td><strong>TO</strong></td>
<td><strong>to</strong></td>
</tr>
<tr>
<td><strong>RB</strong></td>
<td><strong>back</strong></td>
</tr>
<tr>
<td><strong>DT</strong></td>
<td><strong>the</strong></td>
</tr>
<tr>
<td><strong>VB</strong></td>
<td><strong>bill</strong></td>
</tr>
<tr>
<td><strong>NN</strong></td>
<td></td>
</tr>
<tr>
<td><strong>VB</strong></td>
<td></td>
</tr>
<tr>
<td><strong>VBN</strong></td>
<td></td>
</tr>
<tr>
<td><strong>JJ</strong></td>
<td></td>
</tr>
<tr>
<td><strong>NN</strong></td>
<td></td>
</tr>
</tbody>
</table>
Rule-based tagging

1. **Start** with a **dictionary**
2. **Assign all** possible tags to words from the dictionary.
3. **Write rules** (‘by hand’) to selectively **remove** tags
Rule-based tagging example

- Eliminate VBN (past participle) if VBD (past tense) is an option when (VBN|VBD) follows “<s> PRP (personal pronoun)”
- These kinds of rules become *unwieldy* and force determinism where there may not be any.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>she</td>
<td>promised</td>
<td>to</td>
<td>back</td>
</tr>
<tr>
<td>PRP</td>
<td>VBD</td>
<td>TO</td>
<td>RB</td>
</tr>
<tr>
<td>VBN</td>
<td>JJ</td>
<td>NN</td>
<td>DT</td>
</tr>
<tr>
<td>NN</td>
<td>VB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Can we use statistics instead?
Reminder: Bayes’ Rule

\[ P(X, Y) = P(X)P(Y|X) \]
\[ P(X, Y) = P(Y)P(X|Y) \]

\[ P(X|Y) = \frac{P(X)}{P(Y)} P(Y|X) \]
Statistical PoS tagging

- Determine the **most likely** tag sequence \( t_{1:n} \) by:

\[
\arg\max_{t_{1:n}} P(t_{1:n}|w_{1:n}) = \arg\max_{t_{1:n}} \frac{P(w_{1:n}|t_{1:n})P(t_{1:n})}{P(w_{1:n})} \\
= \arg\max_{t_{1:n}} \prod_{i}^{n} P(w_i|t_i)P(t_i|t_{i-1})
\]

By Bayes’ Rule

Only maximize numerator

Assuming independence

Assuming Markov

\( \approx \arg\max_{t_{1:n}} \prod_{i}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \)

**CSC401/2511 – Spring 2017**
Word likelihood probability $P(w_i | t_i)$

- **VBZ** (verb, 3rd person singular present) is likely *is*.
- Compute $P(is|VBZ)$ by counting in a corpus that has already been tagged:

$$P(w_i | t_i) = \frac{\text{Count}(w_i \text{ tagged as } t_i)}{\text{Count}(t_i)}$$

e.g.,

$$P(is|VBZ) = \frac{\text{Count}(is \text{ tagged as } VBZ)}{\text{Count}(VBZ)} = \frac{10,073}{21,627} = 0.47$$
Tag-transition probability $P(t_i | t_{i-1})$

- Will/MD the/DT chair/NN chair/?? the/DT meeting/NN from/IN that/DT chair/NN?

a)

b)
Those are hidden Markov models!

- We’ll see these next week…

Image sort of from *2001: A Space Odyssey* by MGM pictures
Appendix – prepositions from CELEX

<table>
<thead>
<tr>
<th></th>
<th>count</th>
<th></th>
<th>count</th>
<th></th>
<th>count</th>
<th></th>
<th>count</th>
<th></th>
<th>count</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>540,085</td>
<td>through</td>
<td>14,964</td>
<td>worth</td>
<td>1,563</td>
<td>pace</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in</td>
<td>331,235</td>
<td>after</td>
<td>13,670</td>
<td>toward</td>
<td>1,390</td>
<td>nigh</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>for</td>
<td>142,421</td>
<td>between</td>
<td>13,275</td>
<td>plus</td>
<td>750</td>
<td>re</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>125,691</td>
<td>under</td>
<td>9,525</td>
<td>till</td>
<td>686</td>
<td>mid</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with</td>
<td>124,965</td>
<td>per</td>
<td>6,515</td>
<td>amongst</td>
<td>525</td>
<td>o’er</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>on</td>
<td>109,129</td>
<td>among</td>
<td>5,090</td>
<td>via</td>
<td>351</td>
<td>but</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>at</td>
<td>100,169</td>
<td>within</td>
<td>5,030</td>
<td>amid</td>
<td>222</td>
<td>ere</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>by</td>
<td>77,794</td>
<td>towards</td>
<td>4,700</td>
<td>underneath</td>
<td>164</td>
<td>less</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>from</td>
<td>74,843</td>
<td>above</td>
<td>3,056</td>
<td>versus</td>
<td>113</td>
<td>midst</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>about</td>
<td>38,428</td>
<td>near</td>
<td>2,026</td>
<td>amidst</td>
<td>67</td>
<td>o’</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>than</td>
<td>20,210</td>
<td>off</td>
<td>1,695</td>
<td>sans</td>
<td>20</td>
<td>thru</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>over</td>
<td>18,071</td>
<td>past</td>
<td>1,575</td>
<td>circa</td>
<td>14</td>
<td>vice</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Appendix – particles

<table>
<thead>
<tr>
<th>aboard</th>
<th>aside</th>
<th>besides</th>
<th>forward(s)</th>
<th>opposite</th>
<th>through</th>
</tr>
</thead>
<tbody>
<tr>
<td>about</td>
<td>astray</td>
<td>between</td>
<td>home</td>
<td>out</td>
<td>throughout</td>
</tr>
<tr>
<td>above</td>
<td>away</td>
<td>beyond</td>
<td>in</td>
<td>outside</td>
<td>together</td>
</tr>
<tr>
<td>across</td>
<td>back</td>
<td>by</td>
<td>inside</td>
<td>over</td>
<td>under</td>
</tr>
<tr>
<td>ahead</td>
<td>before</td>
<td>close</td>
<td>instead</td>
<td>overhead</td>
<td>underneath</td>
</tr>
<tr>
<td>alongside</td>
<td>behind</td>
<td>down</td>
<td>near</td>
<td>past</td>
<td>up</td>
</tr>
<tr>
<td>apart</td>
<td>below</td>
<td>east, etc.</td>
<td>off</td>
<td>round</td>
<td>within</td>
</tr>
<tr>
<td>around</td>
<td>beneath</td>
<td>eastward(s), etc.</td>
<td>on</td>
<td>since</td>
<td>without</td>
</tr>
</tbody>
</table>
## Appendix – conjunctions

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th></th>
<th>Count</th>
<th></th>
<th>Count</th>
<th></th>
<th>Count</th>
<th></th>
<th>Count</th>
<th></th>
<th>Count</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>and</td>
<td>514,946</td>
<td>yet</td>
<td>5,040</td>
<td></td>
<td>considering</td>
<td>174</td>
<td>forasmuch as</td>
<td>0</td>
<td>however</td>
<td>0</td>
<td>immediately</td>
<td>0</td>
</tr>
<tr>
<td>that</td>
<td>134,773</td>
<td>since</td>
<td>4,843</td>
<td></td>
<td>lest</td>
<td>131</td>
<td>however</td>
<td>0</td>
<td>immediately</td>
<td>0</td>
<td>in as far as</td>
<td>0</td>
</tr>
<tr>
<td>but</td>
<td>96,889</td>
<td>where</td>
<td>3,952</td>
<td></td>
<td>albeit</td>
<td>104</td>
<td>immediately</td>
<td>0</td>
<td>in so far as</td>
<td>0</td>
<td>inasmuch as</td>
<td>0</td>
</tr>
<tr>
<td>or</td>
<td>76,563</td>
<td>nor</td>
<td>3,078</td>
<td></td>
<td>providing</td>
<td>96</td>
<td>inasmuch as</td>
<td>0</td>
<td>insomuch as</td>
<td>0</td>
<td>insomuch that</td>
<td>0</td>
</tr>
<tr>
<td>as</td>
<td>54,608</td>
<td>once</td>
<td>2,826</td>
<td></td>
<td>whereupon</td>
<td>85</td>
<td>like</td>
<td>0</td>
<td>neither nor</td>
<td>0</td>
<td>neither nor</td>
<td>0</td>
</tr>
<tr>
<td>if</td>
<td>53,917</td>
<td>unless</td>
<td>2,205</td>
<td></td>
<td>seeing</td>
<td>63</td>
<td>now</td>
<td>0</td>
<td>now that</td>
<td>0</td>
<td>now that</td>
<td>0</td>
</tr>
<tr>
<td>when</td>
<td>37,975</td>
<td>why</td>
<td>1,333</td>
<td></td>
<td>directly</td>
<td>26</td>
<td>only</td>
<td>0</td>
<td>provided</td>
<td>0</td>
<td>provided</td>
<td>0</td>
</tr>
<tr>
<td>because</td>
<td>23,626</td>
<td>now</td>
<td>1,290</td>
<td></td>
<td>ere</td>
<td>12</td>
<td>provided</td>
<td>0</td>
<td>providing</td>
<td>0</td>
<td>providing</td>
<td>0</td>
</tr>
<tr>
<td>so</td>
<td>12,933</td>
<td>neither</td>
<td>1,120</td>
<td></td>
<td>notwithstanding</td>
<td>3</td>
<td>seeing</td>
<td>0</td>
<td>seeing as</td>
<td>0</td>
<td>seeing as</td>
<td>0</td>
</tr>
<tr>
<td>before</td>
<td>10,720</td>
<td>whenever</td>
<td>913</td>
<td></td>
<td>according as</td>
<td>0</td>
<td>seeing as how</td>
<td>0</td>
<td>seeing that</td>
<td>0</td>
<td>seeing that</td>
<td>0</td>
</tr>
<tr>
<td>though</td>
<td>10,329</td>
<td>whereas</td>
<td>867</td>
<td></td>
<td>as if</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
</tr>
<tr>
<td>than</td>
<td>9,511</td>
<td>except</td>
<td>864</td>
<td></td>
<td>as long as</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
</tr>
<tr>
<td>while</td>
<td>8,144</td>
<td>till</td>
<td>686</td>
<td></td>
<td>as though</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
</tr>
<tr>
<td>after</td>
<td>7,042</td>
<td>provided</td>
<td>594</td>
<td></td>
<td>both and</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
</tr>
<tr>
<td>whether</td>
<td>5,978</td>
<td>whilst</td>
<td>351</td>
<td></td>
<td>but that</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
</tr>
<tr>
<td>for</td>
<td>5,935</td>
<td>suppose</td>
<td>281</td>
<td></td>
<td>but then</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
</tr>
<tr>
<td>although</td>
<td>5,424</td>
<td>cos</td>
<td>188</td>
<td></td>
<td>but then again</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
</tr>
<tr>
<td>until</td>
<td>5,072</td>
<td>supposing</td>
<td>185</td>
<td></td>
<td>either or</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
<td>without</td>
<td>0</td>
</tr>
</tbody>
</table>
### Appendix – Penn TreeBank PoS tags

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordin. conjunction</td>
<td>and, but, or</td>
<td>SYM</td>
<td>symbol</td>
<td>+, %, &amp;</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td>one, two, three</td>
<td>TO</td>
<td>“to”</td>
<td>to</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td>a, the</td>
<td>UH</td>
<td>interjection</td>
<td>ah, oops</td>
</tr>
<tr>
<td>EX</td>
<td>existential ‘there’</td>
<td>there</td>
<td>VBD</td>
<td>verb, past tense</td>
<td>ate</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td>mea culpa</td>
<td>VB</td>
<td>verb, base form</td>
<td>eat</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/sub-conj</td>
<td>of, in, by</td>
<td>VBG</td>
<td>verb, gerund</td>
<td>eating</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td>yellow</td>
<td>VBN</td>
<td>verb, past participle</td>
<td>eaten</td>
</tr>
<tr>
<td>JJR</td>
<td>adj., comparative</td>
<td>bigger</td>
<td>VBP</td>
<td>verb, non-3sg pres</td>
<td>eat</td>
</tr>
<tr>
<td>JJS</td>
<td>adj., superlative</td>
<td>wildest</td>
<td>VBZ</td>
<td>verb, 3sg pres</td>
<td>eats</td>
</tr>
<tr>
<td>LS</td>
<td>list item marker</td>
<td>1, 2, One</td>
<td>WDT</td>
<td>wh-determiner</td>
<td>which, that</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td>can, should</td>
<td>WP</td>
<td>wh-pronoun</td>
<td>what, who</td>
</tr>
<tr>
<td>NN</td>
<td>noun, sing. or mass</td>
<td>llama</td>
<td>WP$</td>
<td>possessive wh-</td>
<td>whose</td>
</tr>
<tr>
<td>NNS</td>
<td>noun, plural</td>
<td>llamas</td>
<td>WRB</td>
<td>wh-adverb</td>
<td>how, where</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, singular</td>
<td>IBM</td>
<td>$</td>
<td>dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td>Carolinas</td>
<td>#</td>
<td>pound sign</td>
<td>#</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
<td>all, both</td>
<td>“</td>
<td>left quote</td>
<td>‘ or “</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>’s</td>
<td>”</td>
<td>right quote</td>
<td>’ or ”</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td>I, you, he</td>
<td>(</td>
<td>left parenthesis</td>
<td>[, (, {, &lt;</td>
</tr>
<tr>
<td>PRPS</td>
<td>possessive pronoun</td>
<td>your, one’s</td>
<td>)</td>
<td>right parenthesis</td>
<td>[), }], &gt;</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td>quickly, never</td>
<td>,</td>
<td>comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td>faster</td>
<td>.</td>
<td>sentence-final punc</td>
<td>. ! ?</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td>fastest</td>
<td>:</td>
<td>mid-sentence punc</td>
<td>; ; ... --</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td>up, off</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>