PartsOfSpeech (POS) Tagging

SUBMITTED BY :

- Tanvi Arora

BROWN Corpus

- The Brown University Standard Corpus of Present-Day American English (or just Brown Corpus) was compiled in the 1960s by Henry Kučera and W. Nelson Francis at Brown University, Providence, Rhode Island as a general corpus (text collection) in the field of corpus linguistics.
- It contains 500 samples of English-language text, totaling roughly one million words, compiled from works published in the United States in 1961.
- The **Brown Corpus** was painstakingly "tagged" with part-of-speech markers over many years.

Longest sentence in Brown corpus is **180** words long….

…… I mean sentences are really that long ?

- **Comprises of 15 categories**
  - Reportage (Politics / Sports / Society / Financial/etc)
  - Reviews (theatre / books / music/ dance)
  - Religion
  - Skill and Hobbies
  - Miscellaneous (Govt Documents / Foundation Reports / Industry Reports / etc)
  - Learned (Natural Sciences / Medicine / Mathematics / etc)
  - Fiction , Fiction:Science , Fiction :Adventure and Western, etc(Novels / short stories)
  - Humor (Novels/Essays)
  - ….

POS Tagging ? …. Also called grammatical tagging

- Last I remember doing POS tagging in High School.

- My problems ..?
  - Long sentences were tough to break
  - Ambiguity

- How many POS tags do you know ?
  - Well I knew just 8…. Until now

What if I let machine do the Tagging ?

- **It can do more than 8 POS tags, definitely**
- **We call them POS Taggers**
  - **ReadyToUse libraries like spaCy, Pattern**
  - **Trainable libraries like nltk's ClassifierBasedPOSTagger or NaiveBayesPOSTagger or RegexPOSTagger**
  - **Developing libraries - Stanfordnlp**
- **I was not perfect and neither are they. So let's evaluate some of them**

Imagine Manually Tagging longest sentence in the Brown Corpus ….

How did I evaluate ?

BROWN

CORPUS

POS

TAGGERS

UNIVERSAL POS Tagset .. smaller known Tagset easy to auto-compare and will be used to compare Brown-Corpus Tags in UNIVERSAL tagset with others in same..

ADJ : Adjective	INTJ : Interjection	PUNCT : punctuation
ADP : AdPosition	NOUN : noun	SCONJ : subordinating conjunction
ADV : Adverb	NUM : numeral	SYM : symbol
AUX : Auxiliary	PART : particle	VERB : verb
CCONJ :coordinating conjunction	PRON : pronoun	X : other
DET : Determiner	PROPN : proper noun	

SO CALLED UNIVERSAL POS Tagset, implemented differently in different POS Taggers and output vary slightly

Pattern.en

The pattern.en module contains a fast part-of-speech tagger for English (identifies nouns, adjectives, verbs, etc. in a sentence), sentiment analysis, tools for English verb conjugation and noun singularization & pluralization, and a WordNet interface.

LONGEST MATCHING SENTENCE : 52 words long

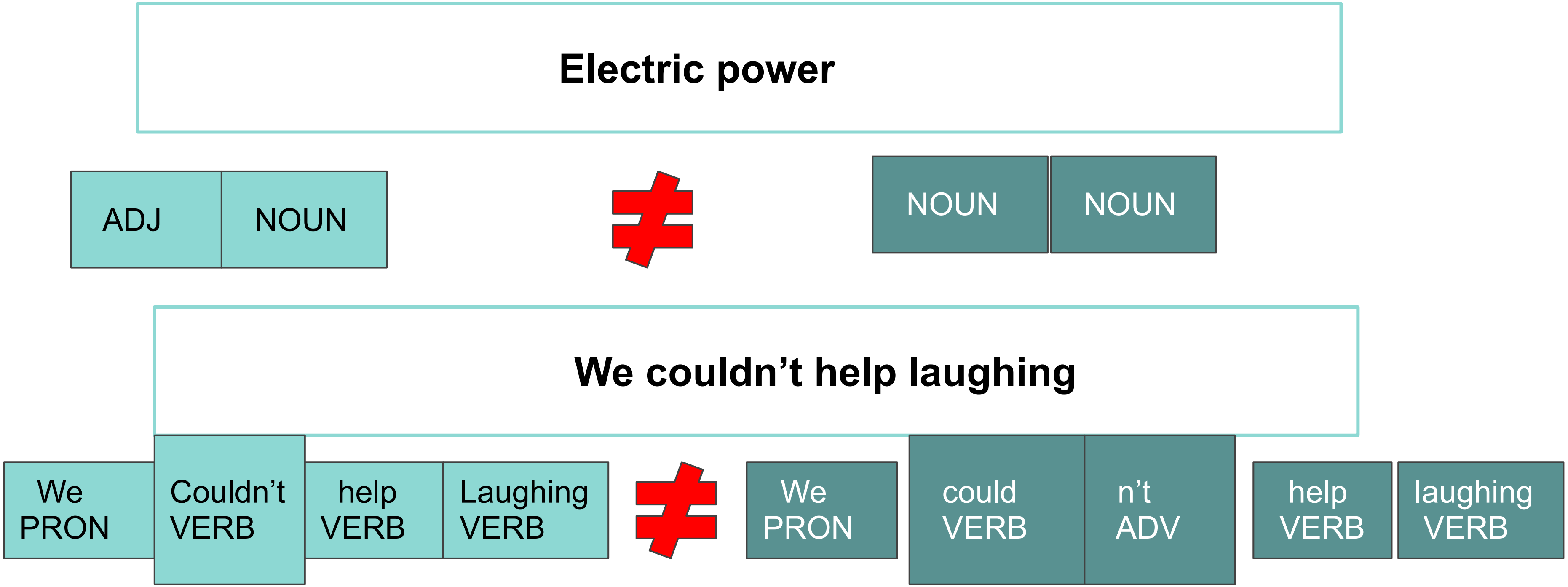
SHORTEST MATCHING SENTENCE : 1 word long

TOTAL MATCHES : 4646 / 57340

Let’s look at couple mismatches from pattern

WHAT DO HUMANISTS SAY ?

WHAT DOES PATTERN SAY ?



Electric power

ADJ	NOUN

≠

NOUN	NOUN

We couldn’t help laughing

We PRON	Couldn’t VERB	help VERB	Laughing VERB

≠

We PRON	could VERB	n’t ADV	help VERB	laughing VERB

spaCy

Is an open-source software library for advanced Natural Language Processing, written in the programming languages Python and Cython. The library is published under the MIT license and currently offers statistical neural network models for English, German, Spanish, Portuguese, French, Italian, Dutch and multi-language NER, as well as tokenization for various other languages

LONGEST MATCHING SENTENCE : 57 words long

SHORTEST MATCHING SENTENCE : 1 word long

TOTAL MATCHES : 8978 / 57340

Let’s look at couple sentences from spaCy

WHAT DO HUMANISTS SAY ?

WHAT DOES SPACY SAY ?

Electric power

ADJ	NOUN	=	ADJ	NOUN

We couldn’t help laughing

We PRON	Couldn’t VERB	help VERB	Laughing VERB	≠	We PRON	could AUX	n’t ADV	help VERB	laughing VERB

BYOM (Nltk- ClassifierBased) Machine learning based

Lets us train a tagger by using a supervised learning algorithm .

We trained our model using Brown Corpus itself.

LONGEST MATCHING SENTENCE : 82 words long

SHORTEST MATCHING SENTENCE : 1 word long

TOTAL MATCHES : 31360 / 57340

Let's look at couple sentences from ClassifierBased

WHAT DO HUMANISTS SAY ?

WHAT DOES SPACY SAY ?

Electric power

ADJ	NOUN

=

ADJ	NOUN

We couldn't help laughing

We PRON	Couldn't VERB	help VERB	Laughing VERB

=

We PRON	couldn't VERB	help VERB	laughing VERB

Which is the BEST ?

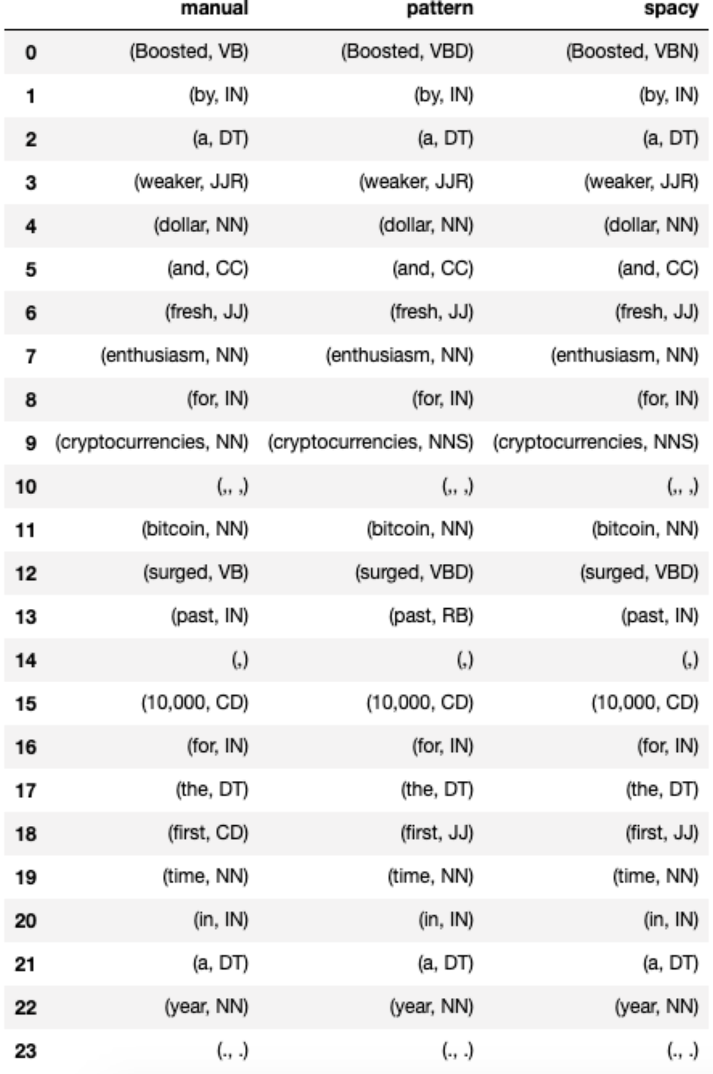
- spaCy clearly had more matches almost double than pattern. It was slow though.

- Each model performed differently, although ClassifierBased POS tagger had maximum matches. But we trained our model using BrownCorpus itself. Wonder what would be the performance if we used some other corpus for training our model ?

Did I compare them with myself ?

YOU BET I DID !!!

- I got 4 words wrong
- Pattern and spacy had a tie. They mismatched on 1 word . I have to find a JUDGE !!



	manual	pattern	spacy
0	(Boosted, VB)	(Boosted, VBD)	(Boosted, VBN)
1	(by, IN)	(by, IN)	(by, IN)
2	(a, DT)	(a, DT)	(a, DT)
3	(weaker, JJR)	(weaker, JJR)	(weaker, JJR)
4	(dollar, NN)	(dollar, NN)	(dollar, NN)
5	(and, CC)	(and, CC)	(and, CC)
6	(fresh, JJ)	(fresh, JJ)	(fresh, JJ)
7	(enthusiasm, NN)	(enthusiasm, NN)	(enthusiasm, NN)
8	(for, IN)	(for, IN)	(for, IN)
9	(cryptocurrencies, NN)	(cryptocurrencies, NNS)	(cryptocurrencies, NNS)
10	(,, ,)	(,, ,)	(,, ,)
11	(bitcoin, NN)	(bitcoin, NN)	(bitcoin, NN)
12	(surged, VB)	(surged, VBD)	(surged, VBD)
13	(past, IN)	(past, RB)	(past, IN)
14	(,)	(,)	(,)
15	(10,000, CD)	(10,000, CD)	(10,000, CD)
16	(for, IN)	(for, IN)	(for, IN)
17	(the, DT)	(the, DT)	(the, DT)
18	(first, CD)	(first, JJ)	(first, JJ)
19	(time, NN)	(time, NN)	(time, NN)
20	(in, IN)	(in, IN)	(in, IN)
21	(a, DT)	(a, DT)	(a, DT)
22	(year, NN)	(year, NN)	(year, NN)
23	(., .)	(., .)	(., .)

THANK YOU