

Story Analyzer:

A Dashboard of NLP Results

Mike Mitri
CIS & Business Analytics Department
College of Business

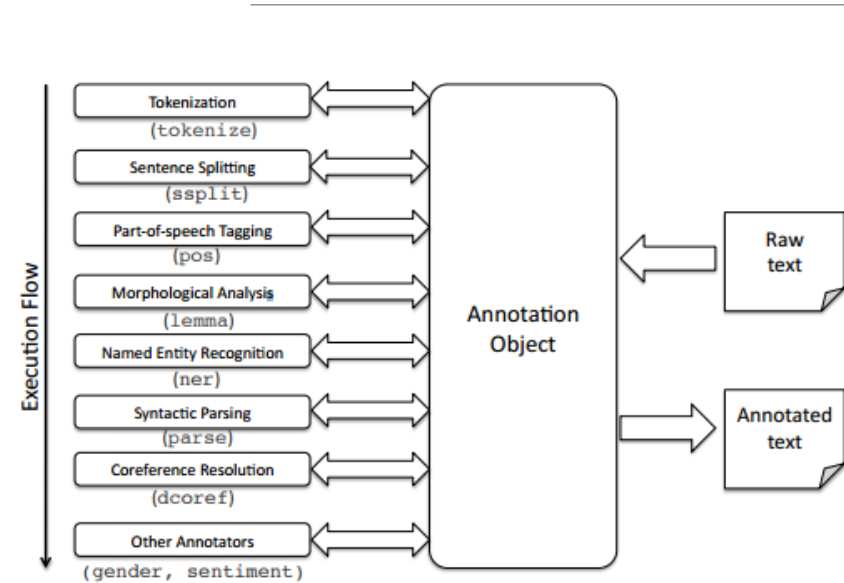


Stanford's CoreNLP

An open source Java-based API of classes and functions that can do several things:

- Breaking a text document into individual sentences
- Tokenizing a sentence (breaking it into individual “words”)
- Identifying parts of speech (POS) within a sentence (nouns, verbs, adjectives, adverbs, etc.)
- Named entity recognition: Recognizing names of people, places, organizations
- Constituency parsing
- Dependency parsing
- Co-reference resolution – finding all expressions that refer to the same entity in a text. (e.g. finding connections nouns and their associated pronouns)
- Temporal tagging – recognizing and normalizing temporal expressions

Stanford CoreNLP Features



Some annotators based on machine learning, others rule-based.

Named Entity Recognition:

1 President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

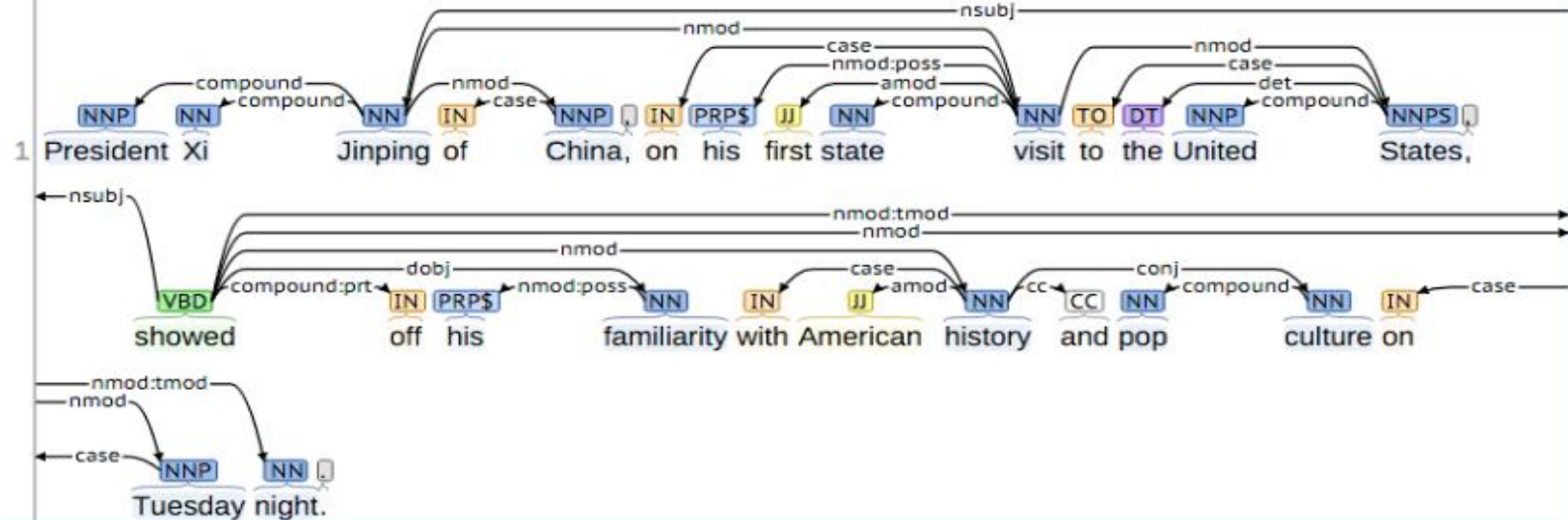
Entities identified: Person (President Xi Jinping), Loc (China), ORDINAL (first), Location (United States), Misc (American), Date (Tuesday), Time (night).

Coreference:

1 President Xi Jinping of China, on his first state visit to the United States, showed off his familiarity with American history and pop culture on Tuesday night.

Coreference links: Mention (President Xi Jinping of China) to M (his).

Basic Dependencies:



Sentence Splitting

Sentences end in periods, question marks, or exclamation points.

But just because you have a period doesn't mean you are at the end of a sentence:

- Mr. Jones
- Samantha G. Jones
- Here is some text (i.e. something written).
- A, B, C, etc., etc., etc.
- Go to website cob.jmu.edu.

Tokenizing

Sentences are made up of words. Tokenizing splits up the sentence into its individual words.

At its simplest, tokenizing uses spaces as delimiters between words.

But, sometimes one word is actually a contraction of two:

- I'm, let's, isn't, won't

CoreNLP tokenizing also splits these into their individual constituents.

POS Tagging

Once tokenized, the individual tokens can be recognized as **parts of speech**.

Stanford's CoreNLP uses the Penn Treebank Tag Set for recognizing parts of speech.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential <i>there</i>
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP\$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	<i>to</i>
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP\$	Possessive wh-pronoun
36.	WRB	Wh-adverb

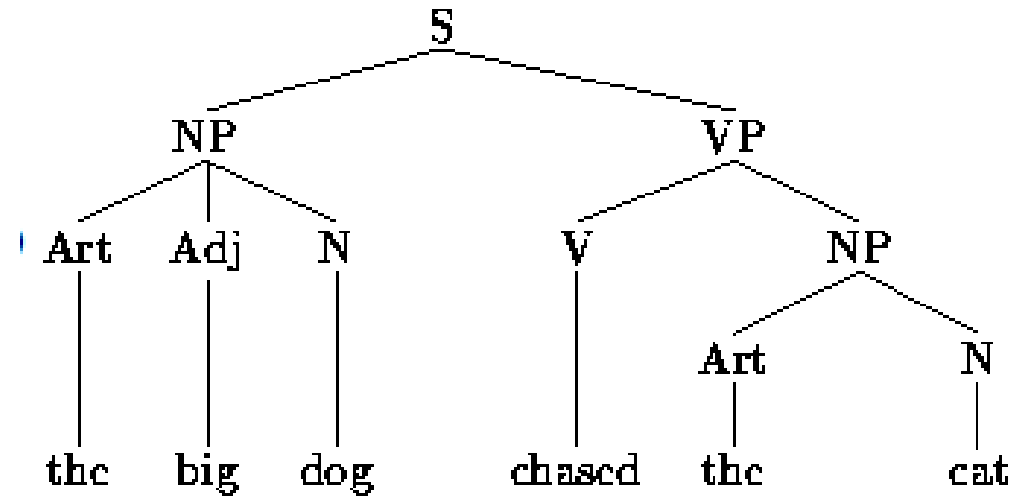
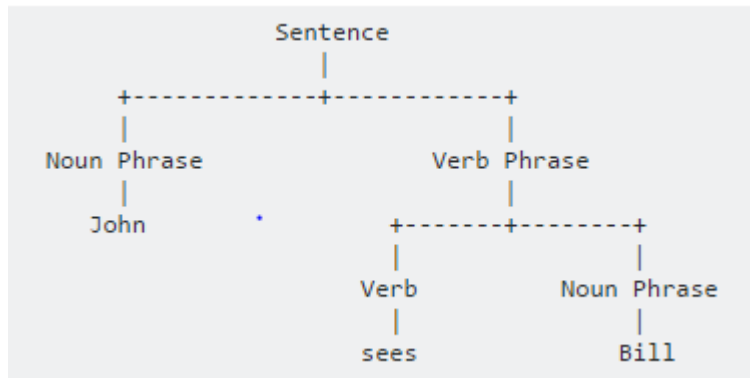
Named Entity Recognition

CoreNLP has models and classes for recognizing the names of:

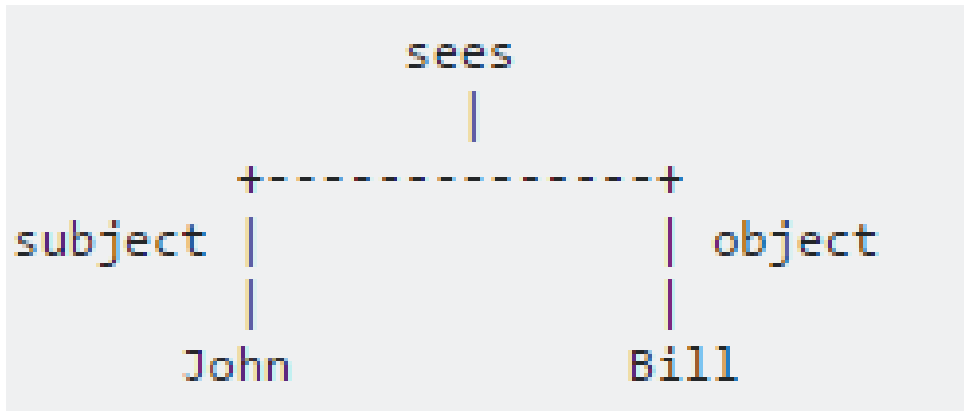
- People
- Places
- Organizations
- Currency
- Time and date
- Extended version: Nationality, Religion, Ideology, Country, State/Province, City, and others.

Constituency parsing

Hierarchy of phrases, sub-phrases, etc.



Dependency parsing



Dependency relationships
(binary predicates) between
words in a sentence.

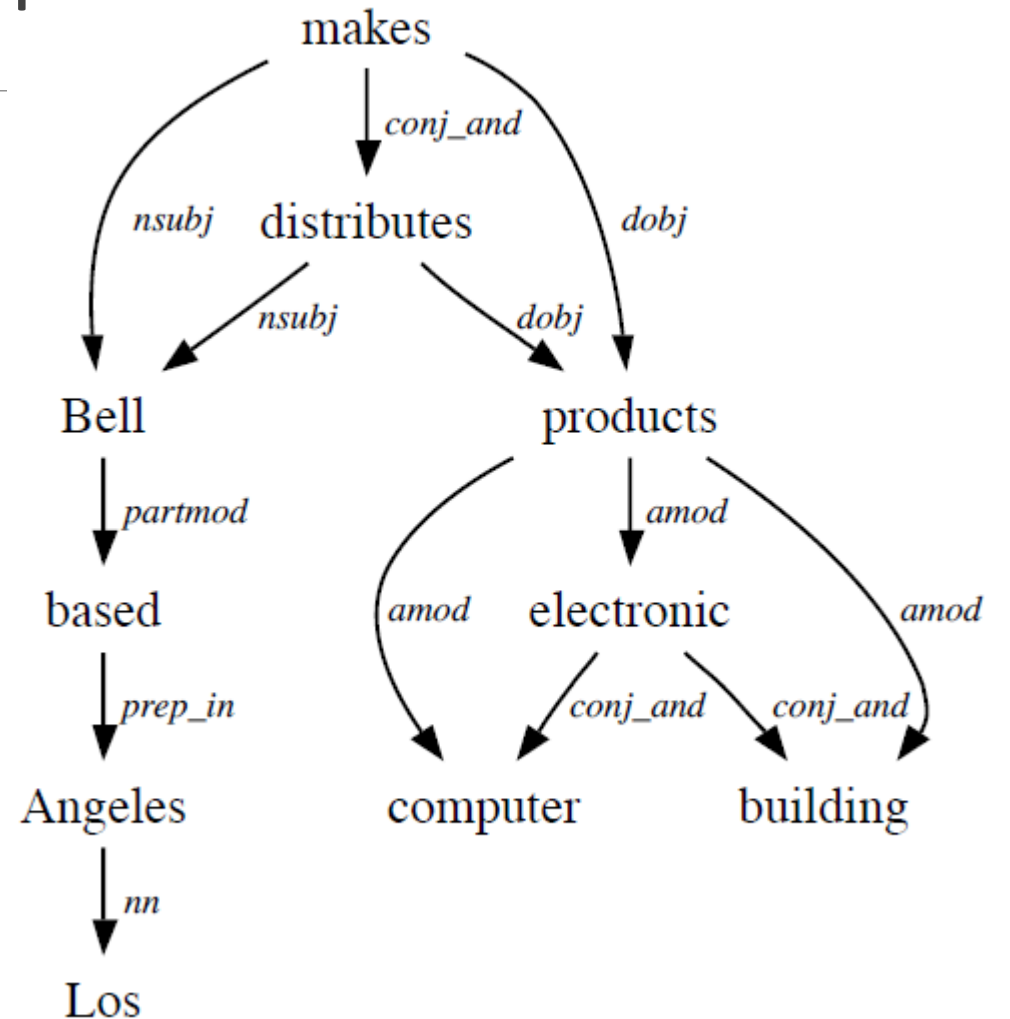
In this picture, the **dependent**
points to the **governor**

Example Dependency Graph

“Bell, based in Los Angeles, makes and distributes electronic, computer and building products.”

From Stanford Typed Dependencies Manual (2008)
http://nlp.stanford.edu/software/dependencies_manual.pdf

Each connecting line is a dependency relationship. In this figure, the **governor** points to the **dependent**.



Dependencies from previous graph

nsubj – nominal subject

nsubjpass – nominal passive subject

dobj – direct object

amod – adjectival modifier

conj_and – conjoint and

prep – preposition (e.g. in, on, at, etc.)

nn – noun compound modifier

partmod – participial verb modifier

How to recognize subject-object relationships in text?

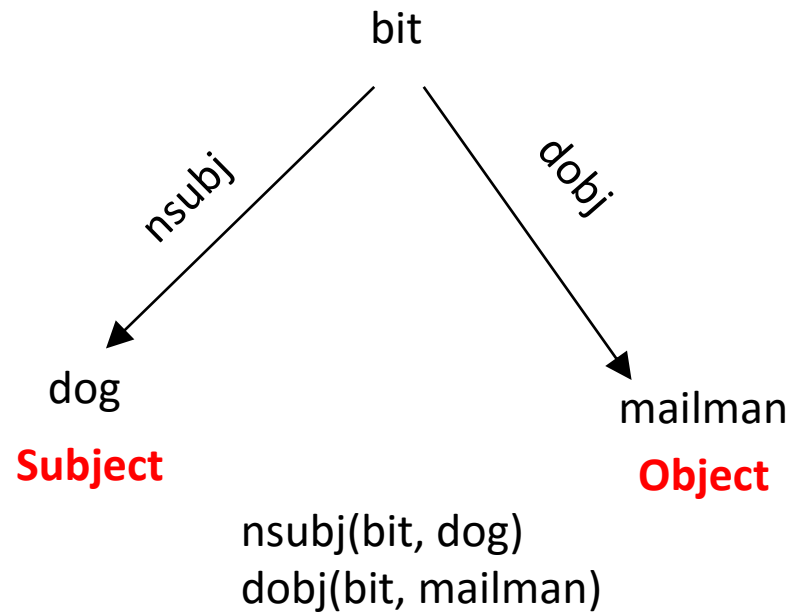
Subject-object relationships -- Who did what to whom?

- The dog bit the mailman
- The mailman was bit by the dog
- Trump beat Rubio in Florida, but he was defeated by Kasich in Ohio.
- The sun exerts gravity on the earth and on Mars.
- Hurricane Matthew bashes Florida with 100mph winds

Subject-object relationships

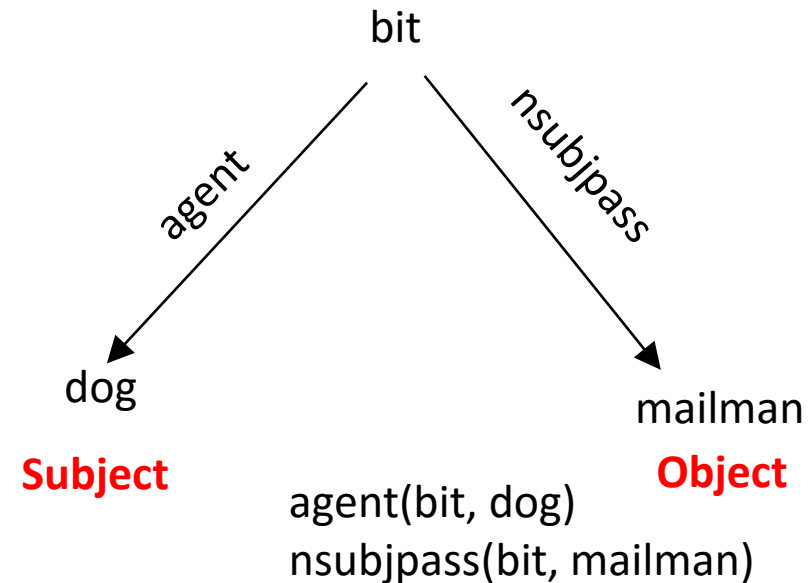
Active Voice

The dog bit the mailman.



Passive Voice

The mailman was bit by the dog.



Key subject-object dependency subgraphs in these sentences

Donald J. Trump beat[a] Marco Rubio in Florida, but then he[a] was defeated by John Kasich[a] in Ohio. Later, though, he[b] beat[b] Kasich[b] and all others at the convention.

- nsubj(beat[a],Trump), dobj(beat[a],Rubio)
- nsubjpass(defeated,he[a]), agent(defeated,Kasich[a])
- nsubj(beat[b],he[b]), dobj(beat[b],Kasich[b])

Note: there are six entities (subjects or objects) cited: Trump, Rubio, he[a], Kasich[a], he[b], and Kasich[b]

Coreference Resolution

Identification of **coreference chains**

A coreference chain has a list of **mentions**

Each mention refers to a word (or cluster of words) in the text

- Mention type – list, nominal, pronominal, proper
- Gender – male, female, or neutral
- Animacy – animate or inanimate

Coreference Resolution

Donald J. Trump beat[a] Marco Rubio in Florida, but then he[a] was defeated by John Kasich[a] in Ohio. Later, though, he[b] beat[b] Kasich[b] and all others at the convention.

What are the **coreference chains**?

1. Trump and he[a] and he[b]
2. Kasich[a] and Kasich[b]

Resulting dependencies after **coreference resolution**. Results in three entities (Trump, Rubio, and Kasich[a]):

- nsubj(beat[a], Trump), dobj(beat[a], Rubio)
- nsubjpass(defeated, Trump), agent(defeated, Kasich[a])
- nsubj(beat[b], Trump), dobj(beat[b], Kasich[a])

Limitations of NLP

Accuracy measures of CoreNLP annotators.

CoreNLP Annotator	F1 Score	Test Information Source
POS Tagging	97	https://nlp.stanford.edu/software/pos-tagger-faq.html
Dependency Parsing	81	https://nlp.stanford.edu/software/stanford-dependencies.shtml
Named Entity Recognition	81	https://nlp.stanford.edu/software/crf-faq.shtml
Coreference Resolution (NN)	60	https://stanfordnlp.github.io/CoreNLP/coref.html

What is an F1 Score?

In machine learning, an F1 score is a metric for measuring accuracy in machine learning (data mining)

Machine learning model will make a prediction. For example prediction of whether a customer will buy a product (simple yes/no). The accuracy of the model is how reliable the prediction is.

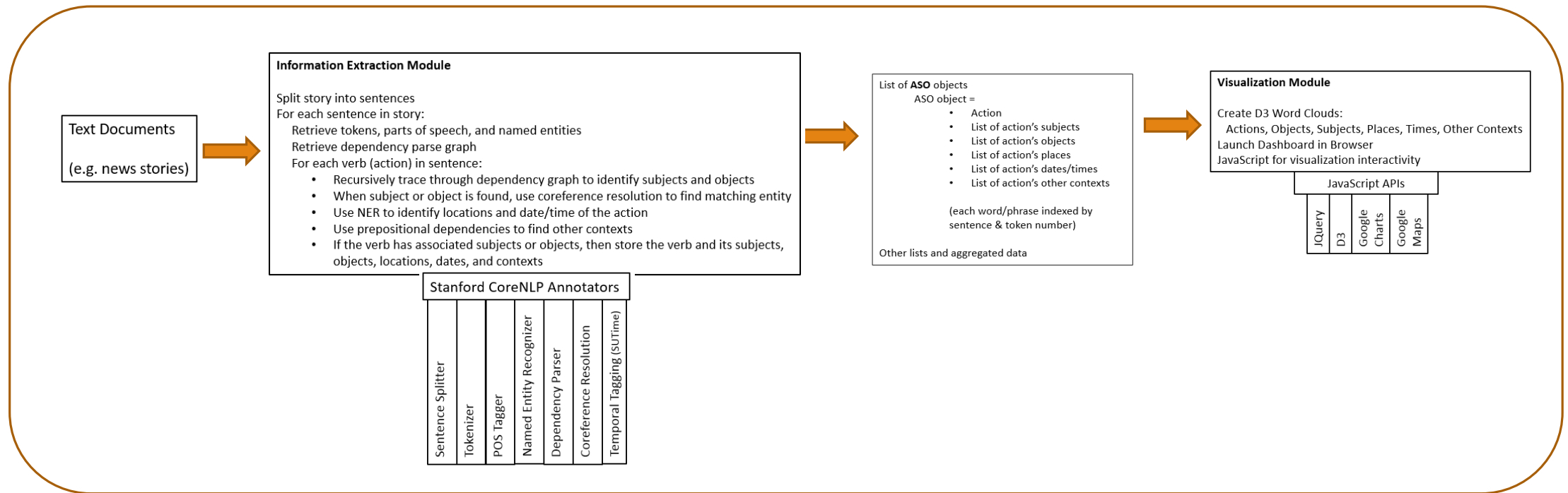
F1 is a score between 0 and 100

Two components of F1:

- **Precision** = What percentage of times did the model predicted yes and the customer actually bought a product.
- **Recall** = What percentage of times that the customer actually bought a product did the model predict this?

There is often a tradeoff between precision and recall. Think about law enforcement.

Story Analyzer's System Architecture



A Story to Visualize (1021 words)

Before summer of 2016, almost nobody expected Donald J. Trump to be our next president. Although he had announced his candidacy the previous summer, many people thought this was just a publicity stunt at the time. Most pundits expected Hillary Clinton to win the presidency. But over time, Trump appealed to a core base of die-hard supporters, and these supporters proved us wrong. Trump stimulated nationalist sentiment by demonizing Islam and rebuking undocumented immigrants. He called for a travel ban against immigrants from Muslim countries and wanted a border wall on the Mexican border. Although many Americans were repulsed by this tactic, others appreciated his "America first" rhetoric.

During the Republican primaries, Trump beat Marco Rubio in Florida and was defeated by John Kasich in Ohio. Throughout the campaign in 2016, he repeatedly referred to Rubio as "Little Marco", and he castigated Ted Cruz as "Lyn' Ted". Contrary to conventional wisdom, Trump beat Kasich, Rubio, Cruz, and all the other presidential candidates at the Republican national convention in Cleveland Ohio in July. Also in July, Clinton defeated Bernie Sanders for the democratic nomination. Ultimately, Trump beat Hillary in the general election on Nov. 8,,although Clinton beat Trump in Virginia and in Maryland. Overall, the popular vote favored Clinton, but the Electoral College chose Trump.

After his inauguration on January 20, 2017, Trump selected several cabinet appointees. He also fought with the press (who he calls the "dishonest media", and he tweeted several times to the American people. On January 30, Trump fired Acting Attorney General Sally Yates after she would not order DOJ employees to enforce President Trump's travel ban due to doubts over its legality. On January 31, Trump nominated Neil Gorsuch to the Supreme Court.

In February, Congress approved many of Trump's cabinet appointees. Also in February, Trump fired Mike Flynn because Flynn had misled vice president Mike Pence in January about his December conversations with Russian officials.

On March 5, Trump tweeted that Barack Obama had wiretapped his offices at Trump Tower. White House officials spent much of the following few weeks attempting to clarify Trump's claim. He continued to fight allegations of collusion with the Russian government. On March 20, FBI director James Comey confirmed that the FBI was indeed investigating the Trump campaign's Russian ties. Ironically, Comey had also derailed Hillary Clinton's campaign by reporting about her emails in October of 2016.

In April, of 2017 the Senate confirmed the Gorsuch nomination, which was a major victory for Trump. Although Trump failed to completely dismantle Obamacare in his first 100 days as he had promised, the House of Representatives narrowly voted to repeal the ACA on May 4. On May 9, Trump fired Comey, stating that this firing relieved unnecessary pressure on the president's ability to engage and negotiate with Russia. On May 24th the Congressional Budget Office estimated that the House health care bill would cause 23 million Americans to lose their insurance. The bill passed but the Senate failed to repeal Obamacare, which is still alive if not kicking. In December of 2017, Congress passed a significant tax cut bill, which was a victory for the president.

On May 17, 2017, Robert Mueller was appointed as a special counsel by Deputy Attorney General Rod Rosenstein. Since then, Mueller has been investigating Trump for possible collusion with Russian power-brokers. After a full year, Mueller hasn't completed his investigation yet, and he is under considerable pressure to do so. Mueller is a Republican and is well respected by Democrats and Republicans alike. Trump is a Republican too, but he believes Mueller is conducting a witch-hunt.

Back in 2006, Trump had a sexual affair with porn star Stormy Daniels. During the campaign in 2016, Trump's personal lawyer Michael Cohen had paid Daniels to refrain from publicizing her affair with Trump. But in early 2018, Stormy began to tell her story, aided by attorney Michael Avenatti. Trump vehemently denies that the affair ever happened. But on July 24th Cohen released an audio recording that suggests otherwise. In this recording, Trump discusses a payoff for Karen McDougal. McDougal had also claimed having an affair with Trump in 2006. Now, in 2018, these dalliances from the past came back to haunt Trump.

Currently, Trump is working with Korean and Chinese leaders, attempting to denuclearize the Korean peninsula. Secretary of State Mike Pompeo actively worked for a summit meeting between Trump and Kim Jong Un. Pompeo's efforts came to fruition at around 10AM on June 12th when Trump and Kim held an historic meeting. Trump had backed out of Obama's nuclear agreement with Iran, insisting that this agreement was a bad deal. So, it will be interesting to see the terms he sets with North Korea. Meanwhile, ISIS is weaker than before but continues to attack Iraqis, Afghanis, and Europeans. Israelis are still fighting Palestinians. Saudi Arabia keeps bombing Yemen.

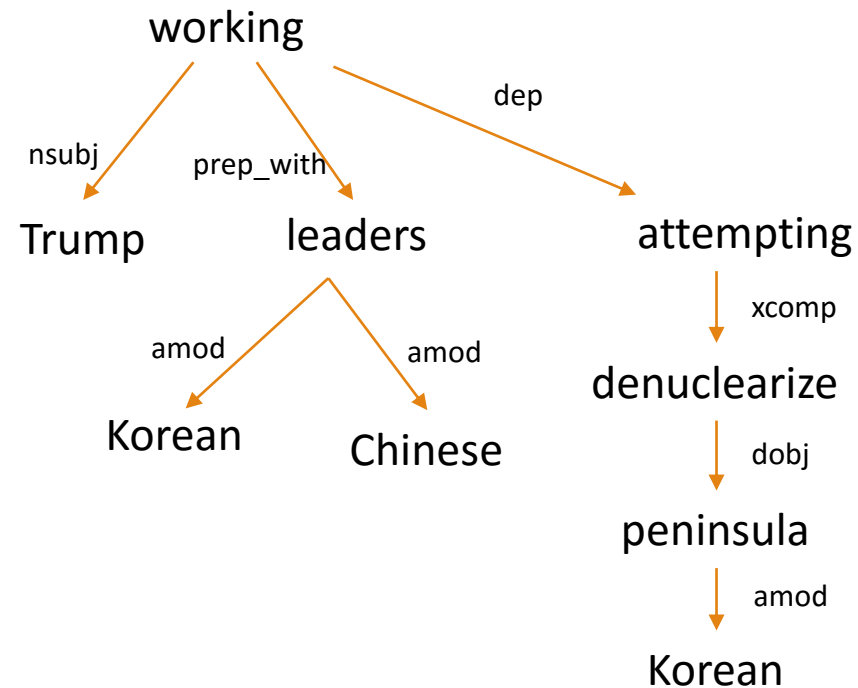
On June 1st, the Trump administration started imposing tariffs against the European Union, Canada, and Mexico. His actions include a 25% tariff on imports of steel and a 10% tariff on aluminum. The tariffs angered U.S. allies, who planned retaliatory tariffs on U.S. goods, and heightened chances of a trade war. The same month, the immigration debate heated up considerably. Attorney General Jeff Sessions ratcheted up the pressure by separating illegal immigrants from their children. Most people see this as cruel and unfair, but some others think it's a necessary part of border control. Trump wants his beautiful wall; perhaps he sees the kids as a bargaining chip.

During the campaign Trump had claimed that NATO is obsolete. This July, he again bashed NATO, claiming that NATO countries are not paying their fair share for defense. NATO allies are also worrying about Trump's friendly overtures to Russia.

The American people elected Donald Trump. He is now the president of the United States. Now "the Donald" is leading us, and we are anxiously watching to see where he takes us. Will peace come to Korea? Will Trump try to fire Mueller? Will Flynn, Cohen, or Paul Manafort eventually flip against Trump? Will we become less divided about immigration? Time will tell.

Currently, Trump is working with Korean and Chinese leaders, attempting to denuclearize the Korean peninsula.

```
root(ROOT-0, working-5)
advmod(working-5, Currently-1)
punct(working-5, ,-2)
nsubj(working-5, Trump-3)
aux(working-5, is-4)
amod(leaders-10, Korean-7)
conj_and(Korean-7, Chinese-9)
amod(leaders-10, Chinese-9)
prep_with(working-5, leaders-10)
punct(working-5, ,-11)
dep(working-5, attempting-12)
aux(denuclearize-14, to-13)
xcomp(attempting-12, denuclearize-14)
det(peninsula-17, the-15)
amod(peninsula-17, Korean-16)
dobj(denuclearize-14, peninsula-17)
punct(working-5, .-18)
```



Coreference chains

COREFERENCE CHAINS

Chain: 192

23 million Americans to => (Sentence 28 Tokens 17-20) PROPER UNKNOWN ANIMATE

their insurance => (Sentence 28 Tokens 22-23) PRONOMINAL UNKNOWN ANIMATE

Chain: 256

an audio recording that suggests otherwise . => (Sentence 40 Tokens 7-13) NOMINAL NEUTRAL INANIMATE

this recording , => (Sentence 41 Tokens 2-4) NOMINAL NEUTRAL INANIMATE

Chain: 129

vice president Mike => (Sentence 19 Tokens 13-15) NOMINAL MALE ANIMATE

his December => (Sentence 19 Tokens 20-21) PRONOMINAL MALE ANIMATE

Chain: 194

the House health care bill would => (Sentence 28 Tokens 10-15) NOMINAL NEUTRAL INANIMATE

The bill passed => (Sentence 29 Tokens 1-3) NOMINAL NEUTRAL INANIMATE

Chain: 258

Karen McDougal . => (Sentence 41 Tokens 10-12) PROPER FEMALE ANIMATE

McDougal had => (Sentence 42 Tokens 1-2) PROPER UNKNOWN ANIMATE

Chain: 322

D3 data visualization API

D3 = Data Driven Documents

A sophisticated JavaScript API see <https://github.com/d3/d3/blob/master/API.md>

Gallery: <https://github.com/d3/d3/wiki/Gallery>

Some examples of D3 visualizations:

- Word Clouds
- Chords
- Time formats
- Force network diagram

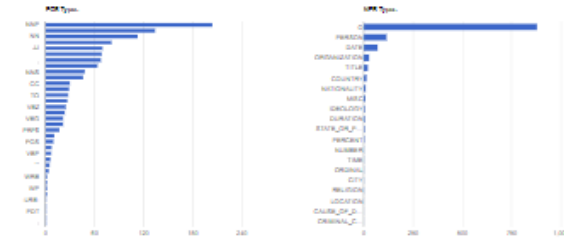
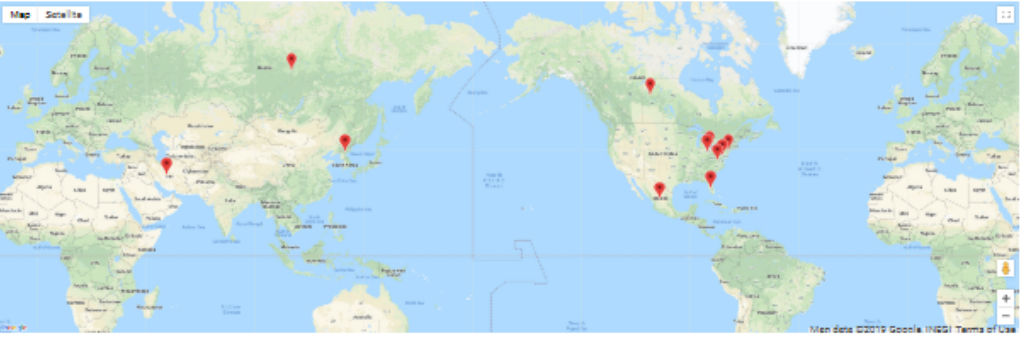
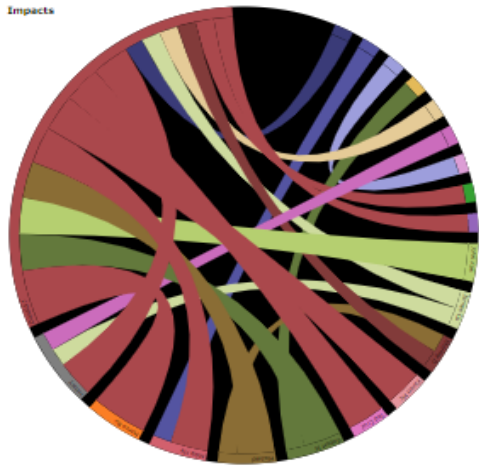
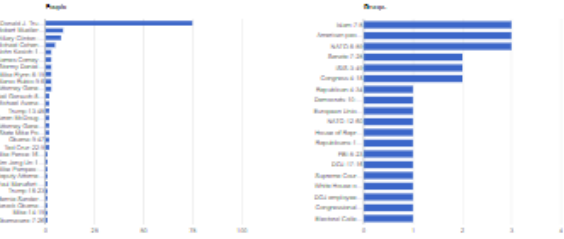
Extremely flexible, but challenging to code

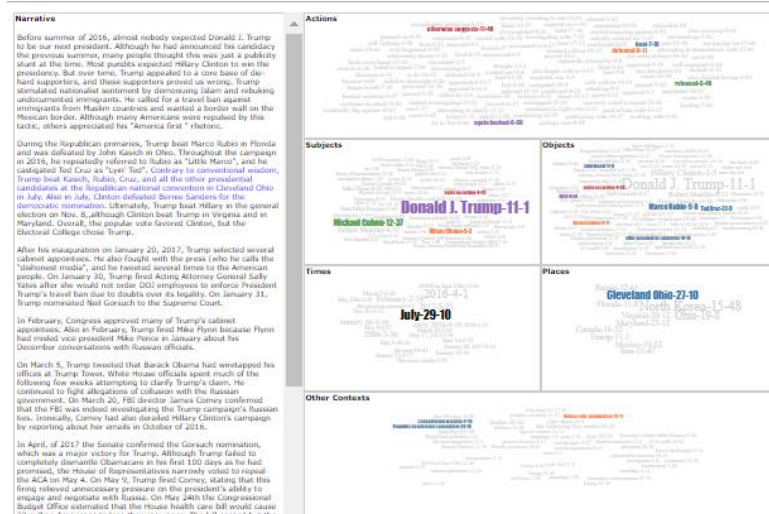
Tableau – easy to use, not particularly flexible

Google charts – moderate difficulty, moderate flexibility

D3 – really flexible, fine-tooth control of individual chart elements. Difficult to program.

Story Analyzer Dashboard

[illegible]

[illegible]

Interactivity:
Selecting an item in
one cloud or
visualization
highlights related
items in other places

