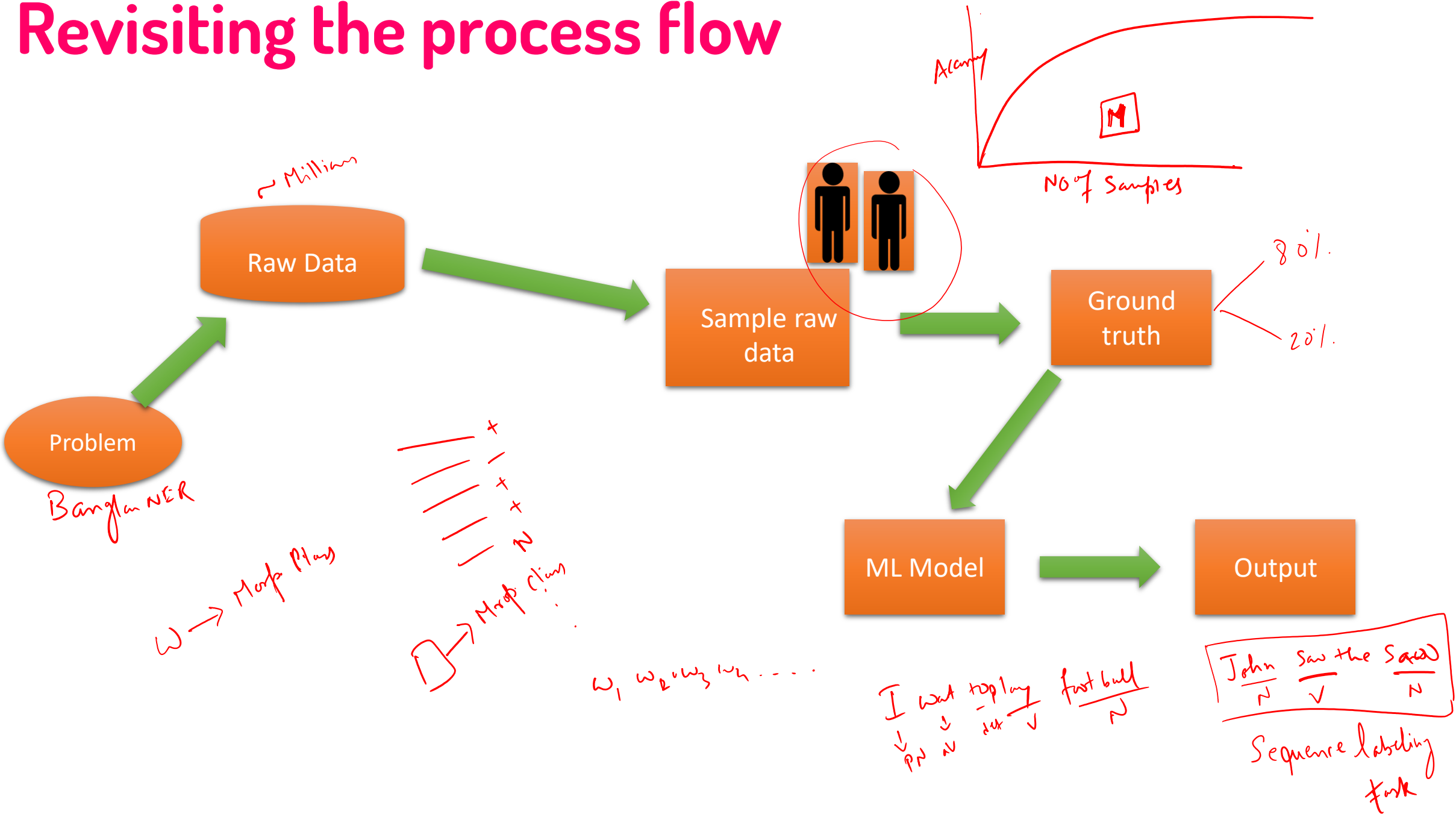


Data Annotation & Evaluation

Part-II

Recap

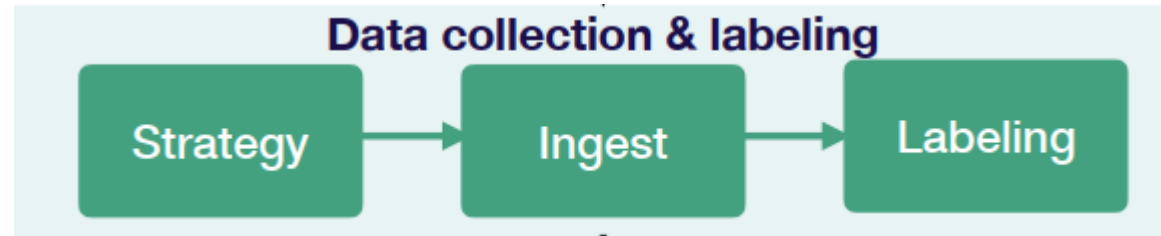
Revisiting the process flow



Data Collection

- After we define what we are going to create, baseline, and metrics in the project, the most painful of the step will begin, **data collection and labeling**.
- Most of Deep Learning applications will require a lot of data which need to be labeled.
 - Time will be mostly consumed in this process.
- Although you can also use public dataset, often that labeled dataset needed for our project is not available publicly.

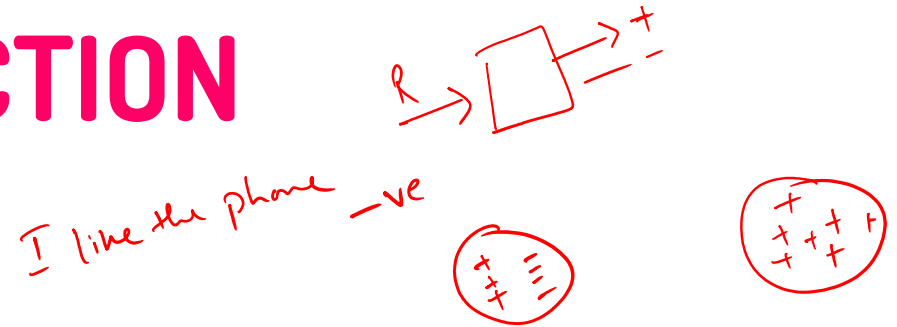
Data Collection



- Strategy:
 - Need to plan how to obtain the complete dataset.
 - Multiple ways to obtain the data
 - The data need to align according to what we want to create in the project.
- Ingest:
 - If the strategy to obtain data is through the internet by scraping and crawling some websites, we need to use some tools to do it.
 - Scrapy is one of the tool that can be helpful for the project

A NOTE ON DATA COLLECTION

- The raw data needs to fit a certain profile.
 - Will also determine how much data needs to be annotated.
- The ideal training corpus has the following key features:
 - It should be **representative**, covering the **domain vocabulary**, **format**, and **genre** of the text.
 - It should be **balanced**, containing instances of each class type that the system is supposed to extract.
 - For example, a system cannot learn to extract corporate entities if there are not enough mentions of corporate entities in the training data.
 - It should be **clean**.
 - It should be **enough**.
 - This ensures accuracy, and is crucial to the creation of a gold standard that will test your system's performance.



DATA ANNOTATION

- Annotation is the act of adding vital information to raw data.
- To a supervised learning algorithm, data without tags is simply noise.
 - Through annotation, however, this noise can be turned into a focused training manual that has an impact all the way up.

Here's an example of some raw text data that could be used to train the entity extractor:

1 Johnny Depp has confirmed his return to the Wizarding World in the new film, Fantastic Beasts: The Crimes of Grindelwald. Best known for playing Captain Jack Sparrow in Pirates of the Caribbean, Depp will star as the eponymous character, dark wizard Gellert Grindelwald. He joins an ensemble cast, also including Eddie Redmayne and Katherine Waterston, for the latest instalment of the popular fantasy series.

1 Person Johnny Depp has confirmed his return to the Wizarding World in the new film, Product Fantastic Beasts: The Crimes of Grindelwald. Best known for playing Title Captain Person Jack Sparrow in Product Pirates of the Caribbean, Per Depp will star as the eponymous character, dark wizard Person Gellert Grindelwald. He joins an ensemble cast, also including Person Eddie Redmayne and Person Katherine Waterston, for the latest instalment of the popular fantasy series.

NER
Stanford NER
SPACY

Example News Text

{ N Person
Loc
org.

Multiple teams of Gandhinagar police's local crime branch (LCB) arrested Shailesh Patel, one of the two prime accused in the Navin Shah kidnapping-murder case, and also seized the SUV in which the murder took place. Shah, 69, was director of printing of Navneet Education and was killed in a moving car on July 25 on SG Road. The LCB officials said that Patel was caught when he went to the spot where he had parked his another car used in the crime on Prantij-Himmatnagar Road. According to the LCB officials, four teams, led by inspector J D Purohit and sub-inspectors H K Solanki, K A Patel and S B Padheriya, are working on the case. After the arrest of two accused — Jignesh Bhavsar and Ramesh Patel — by Ahmedabad city crime branch, LCB seized a car parked by Shailesh Patel outside Devnarayan Dhaba, a roadside eatery, on Monday.

Standard NER Output

Multiple teams of Gandhinagar police's local crime branch (LCB) arrested Shailesh Patel, one of the two prime accused in the Navin Shah kidnapping-murder case, and also seized the SUV in which the murder took place. Shah, 69, was director of printing of Navneet Education and was killed in a moving car on July 25 on SG Road. The LCB officials said that Patel was caught when he went to the spot where he had parked his another car used in the crime on Prantij-Himmatnagar Road. According to the LCB officials, four teams, led by inspector J D Purohit and sub-inspectors H K Solanki, K A Patel and S B Padheriya, are working on the case. After the arrest of two accused — Jignesh Bhavsar and Ramesh Patel — by Ahmedabad city crime branch, LCB seized a car parked by Shailesh Patel outside Devnarayan Dhaba, a roadside eatery, on Monday.

Who/what/where is the

Law enforcement ?

Criminal Name ?

Action taken ?

Victim Name ?

Nature of crime ?

Crime Location ?

CNER

Multiple teams of Gandhinagar police's local crime branch (LCB) arrested Shailesh Patel, one of the two prime accused in the Navin Shah kidnapping-murder case, and also seized the SUV in which the murder took place. Shah, 69, was director of printing of Navneet Education and was killed in a moving car on July 25 on SG Road. The LCB officials said that Patel was caught when he went to the spot where he had parked his another car used in the crime on Prantij-Himmatnagar Road. According to the LCB officials, four teams, led by inspector J D Purohit and sub-inspectors H K Solanki, K A Patel and S B Padheriya, are working on the case. After the arrest of two accused — Jignesh Bhavsar and Ramesh Patel — by Ahmedabad city crime branch, LCB seized a car parked by Shailesh Patel outside Devnarayan Dhaba, a roadside eatery, on Monday.

Law enforcement

Action taken

Nature of crime

Criminal Name

Victim Name

Crime Location

A NOTE ON DATA ANNOTATION

- What is the problem? How complex it is?
- Where is the data coming from?
- Who are the annotators?
- What are the tasks assign to them?

Annotation Tools

- DocAnno
- Brat
- Prodigy
- Tagtog
- DataTurks
- Label Studio
- Stanford Text Annotation Tool

Data Annotation Tool Features

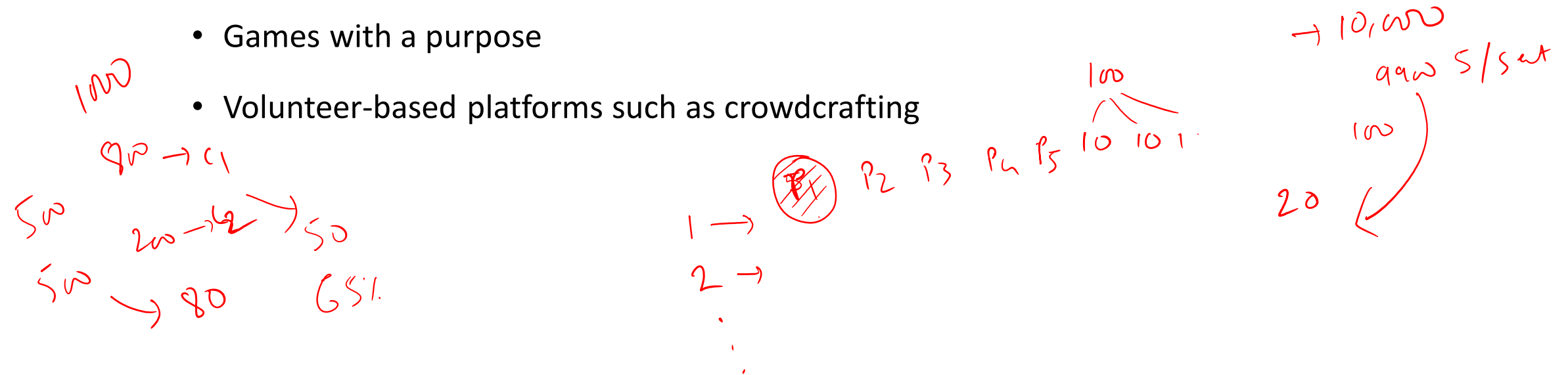
	Pros	Cons
Doccano	web-based, open-source , ability to self-host, Simple, easy-to-use interface, shortcuts for faster labeling, Team collaboration , MIT License	Sometime laggy and unresponsive
Brat	Simple to use, Open-source , easy conversion into other output formats, MIT License, Available API for continuous model training.	Needs to be installed locally, Team collaboration not possible, Outdated UI.
<u>Prodigy</u>	Sleek, modern interface, advance active learning , considerably speeding up the annotation process, Self-hosted, Support for image annotation, Fully scriptable, integrated to spacy, team collaboration .	Expensive
<u>Tagtog</u>	Fast tagging - recognizes all occurrences of an entity once it has been manually labeled and tags them automatically, Support for working with multiple types of data, No installation required, Includes active learning, Team collaboration , Own API for continuous model training.	Expensive , The interface can be a little confusing at first.
DataTurks	Allows Text, image and video labeling, open source , Support for a wide variety of text formats, including PDF, Cloud and local installation, Own API for continuous model training.	Proprietary export format may add friction, support appears to have ceased,
Label Studio	open source , multi-type data, Customizable UI, Quick, easy set up, Mobile friendly	Assets for labeling appear unordered, No simple way of returning to an already labeled asset, No out-of-the-box statistics available

Crowdsourcing

Annotation through Crowdsourcing

"I like the food" → +ve

- Crowdsourcing is an emerging collaborative approach for acquiring annotated corpora and a wide range of other linguistic resources
- Three main kinds of crowdsourcing platforms
 - Paid-for marketplaces such as Amazon Mechanical Turk (AMT) and CrowdFlower (CF)
 - Games with a purpose
 - Volunteer-based platforms such as crowdcrafting



Why Crowdsourcing?

- Paid for crowdsourcing can be 33% **cheaper** than in-house employees when applied to tasks such as tagging and classification (Hoffmann, 2009)
- Games with a purpose can be even cheaper in the long run, since the players are not paid.
- However cost of implementing a game can be higher than AMT/CF costs for smaller projects (Poesio et al, 2012)
- Tap into the **large number of contributors**/players available across the globe, through the internet
- **Easy to reach** native speakers in various languages (but beware Google translate cheaters!)

Genre 1: Mechanized Labor

- Participants (workers) paid a small amount of money to complete easy tasks (HIT = Human Intelligence Task)



Paid for Crowdsourcing

- Contributors are extrinsically motivated through economic incentives
- Carry out micro-tasks in return for micro-payments
- Most NLP projects use crowdsourcing marketplaces: Amazon Mechanical Turk and CrowdFlower
- Requesters post Human Intelligence Tasks (HITs) to a large population of micro-workers (Callison-Burch and Dredze, 2010a)
- Snow et al. (2008) collect event and affect annotations, while Lawson et al. (2010) and Finin et al. (2010) annotate special types of texts such as emails and Twitter feeds, respectively.
- Challenges:
 - **low quality output** due to the workers' purely economic motivation
 - **high costs** for large tasks (Parent and Eskenazi, 2011)
 - **ethical** issues (Fort et al., 2011)

Captcha:










Pick your favorite color:




☒ Red
☐ Green

☐ I'm not a robot  reCAPTCHA

Submit

Select all images with commercial lorries



   [Report a Problem](#)

[Verify](#)

Genre 2: Games with a purpose (GWAPs)

facebook Friends Applications Inbox Home Search

US08 Sentiment Quiz
Play Rankings Awards Feedback Help About

Is the following a **negative**, **neutral** or **positive** statement about the candidate?

“ We are headed down a path that is certain to end in the destruction of our experiment in democracy. ”

− − − + +

Status

4
Level

15:4 13 12

Your current score is **65 points**. Invite your friends and earn 10% of the points they make!

Spread the Word

Tell your Friends! You will earn **10% of your friends' points** after they accept your invitation! The calculation is recursive, so if they invite others you will even get more bonus points.

Sentiment Quiz Award

September 2008
843 players See All

1.	Fiorella	2458
2.	Michel	2241
3.	Birgit	2139
4.	Rose	1011
5.	Herti	930
...		
11.	Arno	101
12.	Guilherme	77
13.	You	65
14.	Lisa	61

Others currently playing

Election Monitor
Vote for your favorite candidate!

Barack Obama John McCain Cynthia McKinney

New Media MBA
www.modul.ac.at/nmt/mba

EDITED BOOK
The Geospatial Web
Geobrowsers, Social Software & the Web 2.0

Page built by Sentiment Quiz (report)

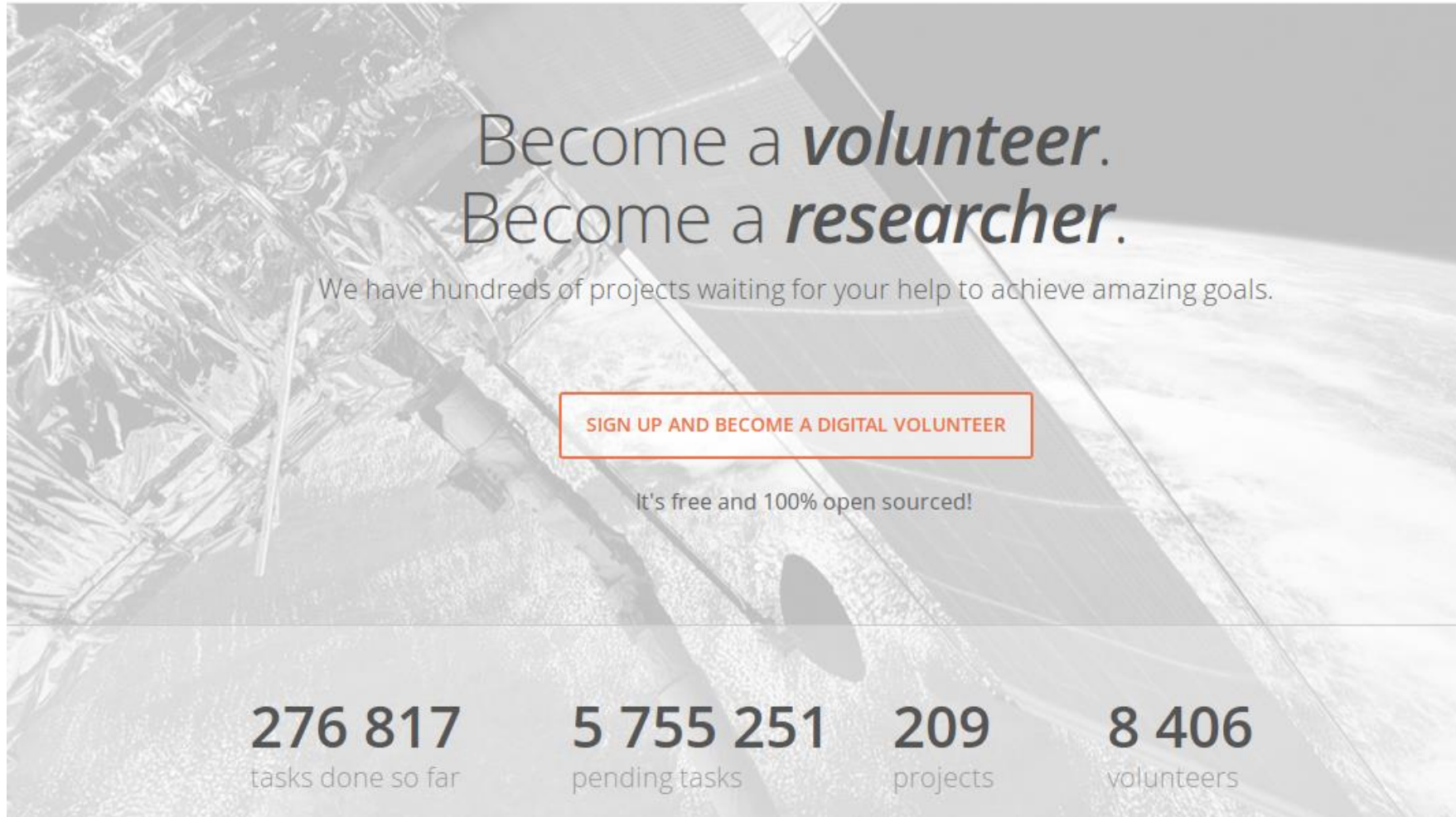

About Find Friends Advertising Developers Terms Privacy Help

?
wordrobe
play what you mean

Ranking (last 50 days)

1	Valerio	32150 points
2	wordrobe	5363 points
3	Aristotle	3998 points
4	sebb	3266 points
5	vincent	3028 points
6	arjanb	2495 points
7	EvaVanmassenhove	1308 points
8	furryfreak	1038 points

Genre 3: Altruistic Crowdsourcing



COMMUNITY PROJECTS ABOUT

SIGN IN [CREATE YOUR PROJECT](#)

Become a ***volunteer.***
Become a ***researcher.***

We have hundreds of projects waiting for your help to achieve amazing goals.

[SIGN UP AND BECOME A DIGITAL VOLUNTEER](#)

It's free and 100% open sourced!

276 817	5 755 251	209	8 406
tasks done so far	pending tasks	projects	volunteers

Task Flow

- Data distribution: how “micro” is each microtask?
 - Long paragraphs hard to digest, worker fatigue
 - Single sentences not always appropriate: e.g. for co-ref
- Reward scheme
 - Granularity – per task? Per set of tasks? High scores?
 - What to do with “bad” work
 - How much to reward
 - No clear, repeatable results for quality: reward relation
 - High rewards get it done faster, but not better
 - Pilot task gives timings, so pay at least minimum wage
- Choose the most appropriate genre or mixture of crowdsourcing genres
 - Trade-offs: Cost; Timescale; Worker skills
- Pilot the design, measure performance, try again
 - Simple, clear design important
 - Binary decision tasks get good results

Evaluation and Corpus Delivery

- Evaluate and aggregate contributor inputs to produce final decision
 - Majority vote
 - Discard inputs from low-trusted contributors (e.g. Hsueh et al. (2009))
 - MACE: a) identify which annotators are trustworthy and b) predict the correct underlying labels (Hovy 2013)
- Merge individual units from the microtasks (e.g. sentences) into complete documents, including all crowdsourced markup
- Tune the expert-created “gold” standard based on annotator feedback
 - Gold standard test questions often contain ambiguities and errors
 - Crowd has a broader knowledge-base than a few experts
 - These are a great opportunity to train workers and amend expert data
 - Better gold data means better output quality, for the same cost
- To facilitate reuse, deliver the corpus in a widely used format, such as XCES, CONLL, GATE XML

Example: CF Instructions

Finding location names in text

Instructions ▲

In each sentence below, mark any names that are locations (e.g. **France**). Don't mark locations that don't have their own name.

There may be no locations in the sentence at all - that's OK.

Examples:

"There was a celebration in **London**"
correct - London is a location name

"The **room** is empty"
wrong, because room isn't the name of a particular location

"We traveled to **Spain** and had a great time **there**"
Only mark the location names, not words that just refer to it

"The award went to **Chelsea** Clinton"
wrong, because here Chelsea is a person

Quality of Annotation

Quality of Annotation

Data types:

- Categorical
 - Binary
 - Sentiment +/-
 - Nominal
 - Hepatitis
 - Viral A, B, C,D, E or auto immune
- Continuous
 - Size of tumor
 - Blood Pressure
 - Quality of answers

How Data are repeated?

- Same input
 - Different observers
 - Inter-rater reliability
 - Same observer at different time
 - Intra-rater reliability

Inter-annotator agreement is a measure of how well two (or more) annotators can make the same annotation decision for a certain category.

Inter-rater reliability: Cohen's Kappa

		Rater 2	
		Correct	Incorrect
Rater 1	Correct	A	B
	Incorrect	C	D

- **A** => The total number of instances that both raters said were correct. The Raters are in agreement.
- **B** => The total number of instances that Rater 2 said was incorrect, but Rater 1 said were correct. This is a disagreement.
- **C** => The total number of instances that Rater 1 said was incorrect, but Rater 2 said were correct. This is also a disagreement.
- **D** => The total number of instances that both Raters said were incorrect. Raters are in agreement.

Inter-rater reliability: Cohen's Kappa

		Rater 2	
		Correct	Incorrect
Rater 1	Correct	A	B
	Incorrect	C	D

$$K = P_o - P_e / 1 - P_e$$

P_o = Number in Agreement / Total

$$P_{(\text{correct})} = (A + B / A + B + C + D) * (A + C / A + B + C + D)$$

$$P_{(\text{incorrect})} = (C + D / A + B + C + D) * (B + D / A + B + C + D)$$

$$P_e = P_{(\text{correct})} + P_{(\text{incorrect})}$$

$K=1 \rightarrow$ Full Agreement
 $K=0 \rightarrow$ Random
 $K<0 \rightarrow$ No effective agreement

p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category.

Text Classification

Positive or negative movie review?

- + *...zany characters and richly applied satire, and some great plot twists*
- *It was pathetic. The worst part about it was the boxing scenes...*
- + *...awesome caramel sauce and sweet toasty almonds. I love this place!*
- *...awful pizza and ridiculously overpriced...*

Positive or negative movie review?

- + ...zany characters and **richly** applied satire, and some **great** plot twists
- It was **pathetic**. The **worst** part about it was the boxing scenes...
- + ...**awesome** caramel sauce and sweet toasty almonds. I **love** this place!
- ...**awful** pizza and **ridiculously** overpriced...

Illustration: Text Classification

- **Supervised Machine Learning**

- A document d
- A fixed set of classes $C = \{c_1, c_2, \dots, c_n\}$
- A training set of m hand-labeled documents $(d_1, c_1), \dots, (d_m, c_k)$.

- Output:

- A learned Classifier function $\varphi: d \rightarrow c$

- Classifiers:

- Naïve Bayes
- Logistic regression
- SVM
- K-NN
- NN
- LSTM $(d) \rightarrow c$
- CNN
- BERT
- ...

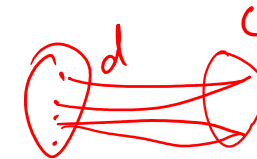
$\begin{matrix} + & - & New \\ c_1 & c_2 & c_3 \end{matrix}$

$\oplus \oplus \ominus \ominus \oplus$ New

0 - neg
10 → pos

0 - S

$P(A|B)$



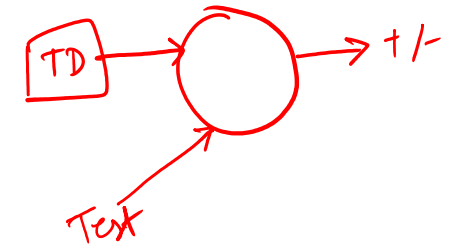
Sample Labelled/Annotated Data

- The script is good.
- The story has satirical humor.
- Too sweet the female character
- Great dialogues!
- Full of adventures and drama
- It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre.
- The greatest screwed up movie !
- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

Sample Labelled/Annotated Data

- The script is good. → +
- The story has satirical humor. → +
- Too sweet the female character → +
- Great dialogues! → +
- Full of adventures and drama → -
- It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. → -
- The greatest screwed up movie ! → -
- unbelievably disappointing → -
- Full of zany characters and richly applied satire, and some great plot twists → +
- this is the greatest screwball comedy ever filmed → +
- It was pathetic. The worst part about it was the boxing scenes. → -

Bag-of-words model
weights
 $P(+ | \text{The})$
 $P(+ | \text{script})$
 $P(+ | \text{good}) \rightarrow \text{high}$
 $P(- | \text{the})$
 $P(- | \text{good}) \dots$
 $P(- | \text{poor}) \uparrow$



ML Model Intuition

- Relies on very simple representation of document
 - Bag of words

$$\varphi \left(\begin{array}{l} \text{I love this movie! It's sweet, but with satirical humor. The} \\ \text{dialogue is great and the adventure scenes are fun... It manages} \\ \text{to be whimsical and romantic while laughing at the conventions} \\ \text{of the fairy tale genre. I would recommend it to just about} \\ \text{anyone. I've seen it several times, and I'm always happy to see it} \\ \text{again whenever I have a friend who hasn't seen it yet.} \end{array} \right) = C$$

Sample Labelled/Annotated Data

- The script is good. → +
- The story has satirical humor. → +
- Too sweet the female character → +
- Great dialogues! → +
- Full of adventures and drama → -
- It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. → -
- The greatest screwed up movie ! → -
- unbelievably disappointing → -
- Full of zany characters and richly applied satire, and some great plot twists → +
- this is the greatest screwball comedy ever filmed → +
- It was pathetic. The worst part about it was the boxing scenes. → -

Sample Labelled/Annotated Data

- The script is good. → +
- The story has satirical humor. → +
- Too sweet the female character → +
- Great dialogues! → +
- Full of adventures and drama → -
- It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. → -
- The greatest screwed up movie ! → -
- unbelievably disappointing → -
- Full of zany characters and richly applied satire, and some great plot twists → +
- this is the greatest screwball comedy ever filmed → +
- It was pathetic. The worst part about it was the boxing scenes. → -

She left me for "good" — -v

→ blood pressure is high — -ve
battery life is high — +ve

+ve
f -ve +ve

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words

$$\varphi \left(\begin{array}{l} \text{I love this movie! It's sweet but with satirical humor. The} \\ \text{dialogue is great and the adventure scenes are fun... It manages} \\ \text{to be whimsical and romantic while laughing at the conventions} \\ \text{of the fairy tale genre. I would recommend it to just about} \\ \text{anyone. I've seen it several times, and I'm always happy to see it} \\ \text{again whenever I have a friend who hasn't seen it yet.} \end{array} \right) = C^{+c_1 - c_2}$$

$$\rightarrow P(c_1 | I) + P(c_1 | \text{love}) + P(c_1 | \text{this}) + P(c_1 | \text{movie!}) + P(c_1 | \text{It's}) + P(c_1 | \text{sweet,}) + P(c_1 | \text{but}) \dots$$

$$\sum P(c_1)$$

$$P(c_2 | I) + P(c_2 | \text{love}) + P(c_2 | \text{this}) + P(c_2 | \text{movie!}) + P(c_2 | \text{It's}) + P(c_2 | \text{sweet,}) + P(c_2 | \text{but}) \dots$$

$$\sum P(c_2)$$

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure ^{happy} scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy I to see it again whenever I have a friend who hasn't seen it yet!

15



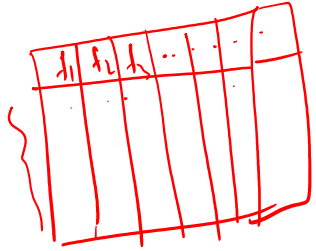
no ordering of words

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

$$P(w_i) = \frac{\text{Count}(w_i)}{\text{Total no. of words}}$$

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words – Which words?



1	1	1	...	1	...

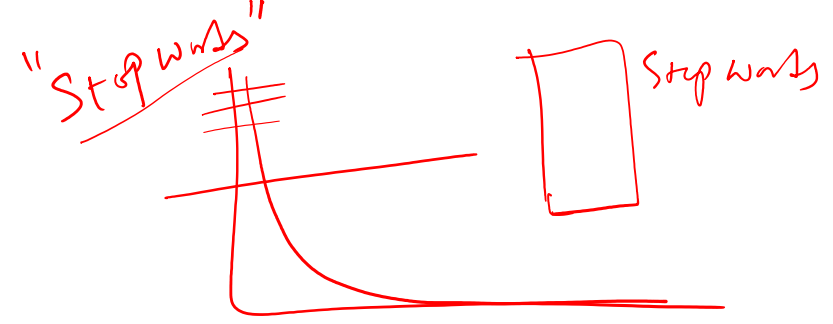
Stop words
Content words (more, sweet)
Out, functional words (the, is, am, we...)

φ (

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

) = C

Books on Tagore
Books of Tagore



ML Model Intuition

- Relies on very simple representation of document
 - Bag of words – Which words?

Word importance

$$\varphi \left(\begin{array}{l} \text{I love this movie! It's sweet, but with satirical humor. The} \\ \text{dialogue is great and the adventure scenes are fun... It manages} \\ \text{to be whimsical and romantic while laughing at the conventions} \\ \text{of the fairy tale genre. I would recommend it to just about} \\ \text{anyone. I've seen it several times, and I'm always happy to see it} \\ \text{again whenever I have a friend who hasn't seen it yet.} \end{array} \right) = C$$

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words – Which words?

$$\varphi \left(\begin{array}{l} \text{I love this movie! It's sweet, but with satirical humor. The} \\ \text{dialogue is great and the adventure scenes are fun... It manages} \\ \text{to be whimsical and romantic while laughing at the conventions} \\ \text{of the fairy tale genre. I would recommend it to just about} \\ \text{anyone. I've seen it several times, and I'm always happy to see it} \\ \text{again whenever I have a friend who hasn't seen it yet.} \end{array} \right) = C$$

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words – Which words?


$$\varphi \left(\begin{array}{l} \text{Love} \\ \text{sweet} \\ \text{satirical} \\ \text{laughing} \\ \text{recommend} \\ \text{several} \\ \text{Times} \\ \text{happy} \\ \text{again} \\ \text{humor} \\ \text{great} \\ \text{adventure} \\ \text{Fun} \\ \text{whimsical} \\ \text{romantic} \end{array} \right) = C$$

ML Model Intuition

- Relies on very simple representation of document
 - Bag of words – Which words?

$$\varphi \left(\begin{array}{l} \text{Love} \\ \text{sweet} \\ \text{satirical} \\ \text{laughing} \\ \text{recommend} \\ \text{several} \\ \text{Times} \\ \text{happy} \\ \text{again} \\ \text{humor} \\ \text{great} \\ \text{adventure} \\ \text{Fun} \\ \text{whimsical} \\ \text{romantic} \end{array} \begin{array}{l} 2 \\ 1 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ \dots \\ \dots \end{array} \right) = \mathcal{C}$$

Frequency/ TF-IDF



Which words?

- Adjectives convey much of the subjective content in a text,
 - a great deal of effort has been devoted extracting semantic orientation (i.e., positive and negative values) for adjectives.
 - *excellent* and *X* → predicts that *X* will be a *positive* adjective, in a situation where we do not know the polarity of X.
 - However, a great deal of sentiment is conveyed through other parts of speech,
 - nouns (masterpiece, disaster),
 - verbs (love, hate),
 - adverbs (skillfully, poorly)

Intensification and Downtoning

- Really *good*
- Very *impressive*
- Arguably *good*
- Somewhat *better* +
- Mostly

+ + + - - -

- (1) I thought this movie would be as good as the Grinch.
- (2) Couldn't you find a more suitable ending? *-ve*

Negation

- (3c) Our Sony phones died after 7 years... which I don't think it's too bad for a cordless phone.
- (3d) I had stayed at Westin hotels before, and was never disappointed until now.
- (3e) Propaganda doesn't succeed because it is manipulative, it works because people WANT it, NEED it, it gives their life a direction and meaning and guards against change. (*The Last Psychiatrist* 2013)

Identifying syntactic patterns

- The food is good → +
- The pasta is awesome → +
- The dessert is too good → +
- The fish is overwhelming → +
- The mutton is amazing → +
- ...
- ...

Identifying syntactic patterns

- The /DT food /NN is /VBZ good /JJ → +
- The /DT pasta /NN is /VBZ awesome /JJ → +
- The /DT dessert /NN is /VBZ too /RB good /JJ → +
- The /DT fish /NN is /VBZ overwhelming /JJ → +
- The /DT mutton /NN is /VBZ amazing /JJ → +
- ...
- ...

Identifying syntactic patterns

- The /DT food /NN is /VBZ good /JJ → +
- The /DT pasta /NN is /VBZ awesome /JJ → +
- The /DT dessert /NN is /VBZ too /RB good /JJ → +
- The /DT fish /NN is /VBZ overwhelming /JJ → +
- The /DT mutton /NN is /VBZ amazing /JJ → +
- ...
- ...

We learn that...

- /DT /NN /VBZ /JJ → +

- That sandwich was delicious → ?
DT NN VBZ JJ

DT NN VBZ JJ → +

Discourse

- (9) It could have been a great movie. It could have been excellent, and to all the people who have forgotten about the older, greater movies before it, will think that as well. It does have beautiful scenery, some of the best since Lord of the Rings. The acting is well done, and I really liked the son of the leader of the Samurai. He was a likeable chap, and I hated to see him die. But, other than all that, this movie is nothing more than hidden rip-offs.

Features

- **Words**
- N-Grams
 - Succeeding words
 - Preceding words
- Skip grams
- Important words
 - Tf-IDF
 - PMI
- POS
 - Frequency
 - N-grams
 - Important POS
 - TF-IDF
 - PMI
- Dependency relations

