KI-Campus
Die Lernplattform
für Künstliche Intelligenz

Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Text Corpus

Salar Mohtaj | DFKI

Text corpus

- What is a text corpus?
- Sample text corpora
- Corpus annotation
- Underfitting and overfitting
- Data splitting
- Text corpora in Python

Text corpus

- What is a text corpus?
- Sample text corpora
- Corpus annotation
- Underfitting and overfitting
- Data splitting
- Text corpora in Python

What is a text corpus?

- Text corpus is a collection of text, usually contains several documents
  - Wikipedia articles
  - Collection of movies reviews
  - Internet comments
  - Collection of tweets
- A corpus may be quite small, for example, containing only **thousands words** of text, or very large, containing **millions of words**
- NLP corpora are in different standards based on the target task

What is a text corpus?

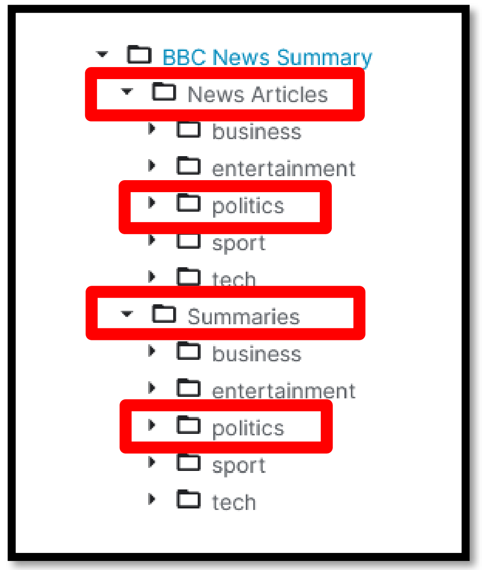
- Text corpora can be compared based on different factors
  - **Size**: larger corpora are better for training deep learning models
  - **Domain**: unless corpus has been collected for specific tasks, it should include different genres such as newspapers, magazines, blogs, academic journals, etc
  - **Metadata**: metadata should indicate the sources, assumptions, limitations and what's included in the corpus
  - **Clean**: a wordlist giving word forms of the same word can be messy to process

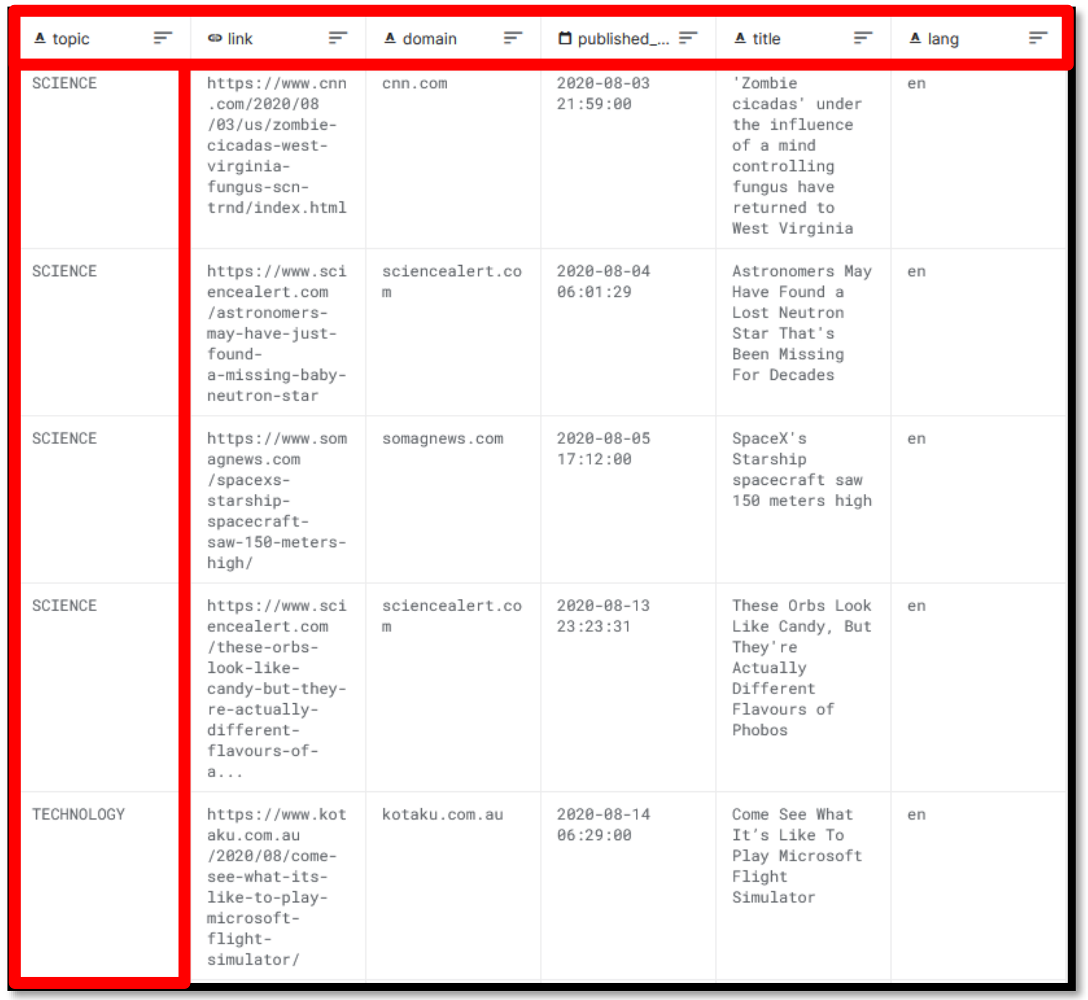
Text corpus

- What is a text corpus?
- Sample text corpora
- Corpus annotation
- Underfitting and overfitting
- Data splitting
- Text corpora in Python

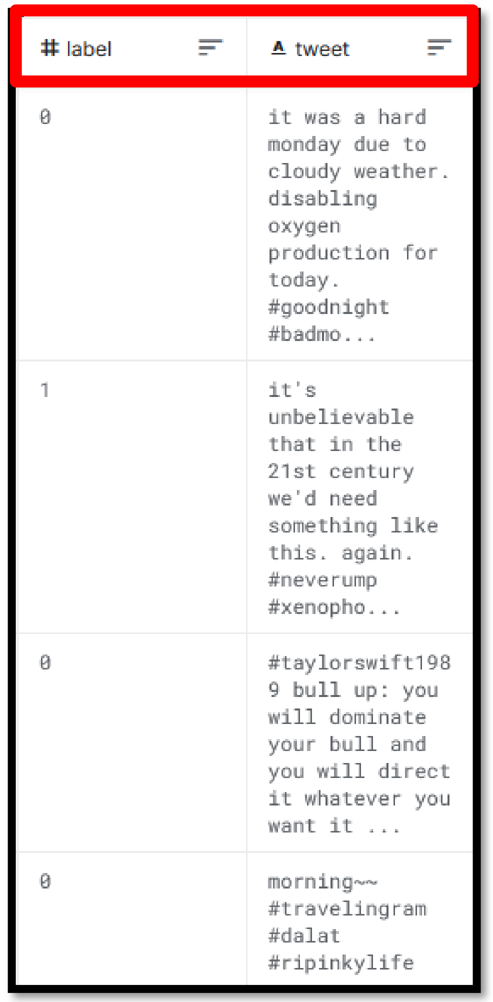
Sample text corpora

- Common file formats
  - TXT
  - CSV
  - JSON
  - XML

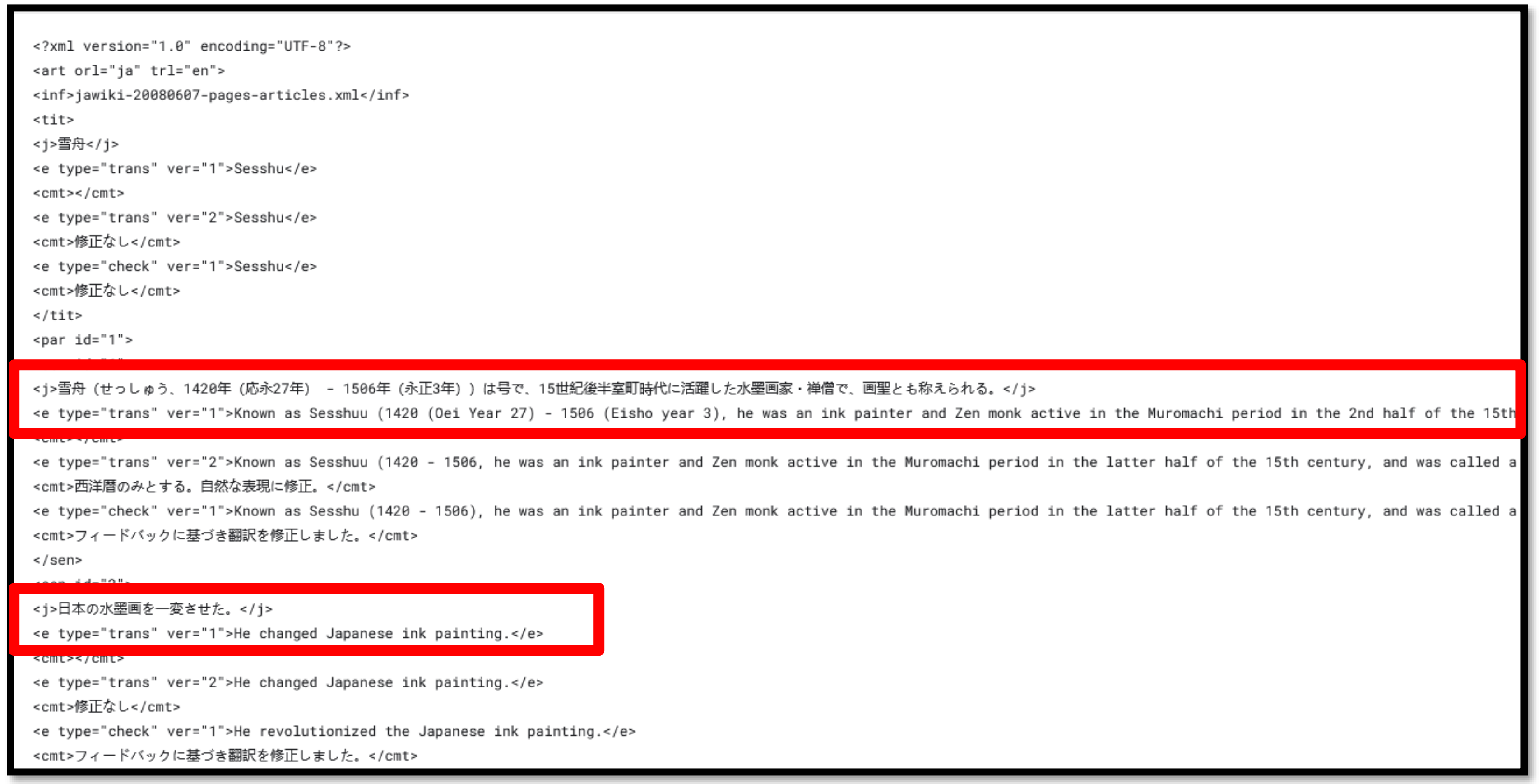




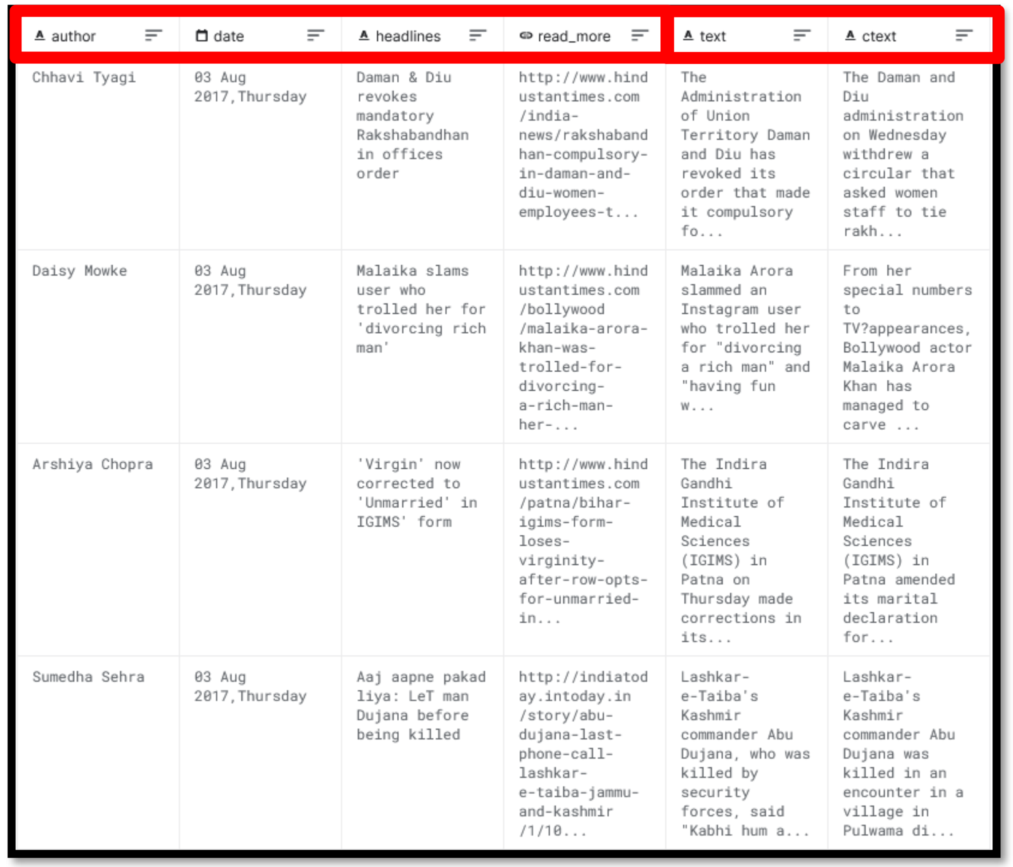
topic	link	domain	published_...	title	lang
SCIENCE	https://www.cnn.com/2020/08/03/us/zombie-cicadas-west-virginia-fungus-scn-trnd/index.html	cnn.com	2020-08-03 21:59:00	'Zombie cicadas' under the influence of a mind controlling fungus have returned to West Virginia	en
SCIENCE	https://www.sciencealert.com/astronomers-may-have-just-found-a-missing-baby-neutron-star	sciencealert.com	2020-08-04 06:01:29	Astronomers May Have Found a Lost Neutron Star That's Been Missing For Decades	en
SCIENCE	https://www.somagnews.com/spacexs-starship-spacecraft-saw-150-meters-high/	somagnews.com	2020-08-05 17:12:00	SpaceX's Starship spacecraft saw 150 meters high	en
SCIENCE	https://www.sciencealert.com/these-orbs-look-like-candy-but-they-re-actually-different-flavours-of-a...	sciencealert.com	2020-08-13 23:23:31	These Orbs Look Like Candy, But They're Actually Different Flavours of Phobos	en
TECHNOLOGY	https://www.kotaku.com.au/2020/08/come-see-what-its-like-to-play-microsoft-flight-simulator/	kotaku.com.au	2020-08-14 06:29:00	Come See What It's Like To Play Microsoft Flight Simulator	en



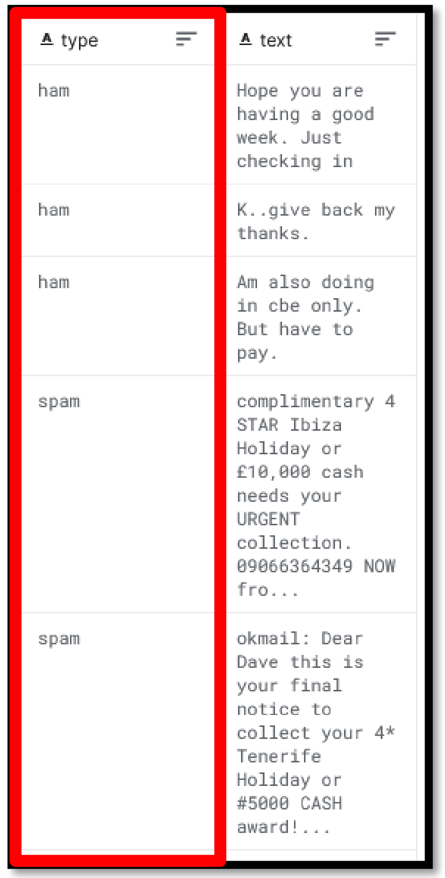
# label	tweet
0	it was a hard monday due to cloudy weather. disabling oxygen production for today. #goodnight #badmo...
1	it's unbelievable that in the 21st century we'd need something like this. again. #neverump #xenopho...
0	#taylorswift1989 bull up: you will dominate your bull and you will direct it whatever you want it ...
0	morning~~ #travelingram #dalat #ripinkylife



```xml
<?xml version="1.0" encoding="UTF-8"?>
<art orl="ja" trl="en">
<inf>jawiki-20080607-pages-articles.xml</inf>
<tit>
<j>雪舟</j>
<e type="trans" ver="1">Sesshu</e>
<cmt></cmt>
<e type="trans" ver="2">Sesshu</e>
<cmt>修正なし</cmt>
<e type="check" ver="1">Sesshu</e>
<cmt>修正なし</cmt>
</tit>
<par id="1">
<sen id="1">
<j>雪舟（せっしゅう、1420年（応永27年） - 1506年（永正3年））は号で、15世紀後半室町時代に活躍した水墨画家・禅僧で、画聖とも称えられる。</j>
<e type="trans" ver="1">Known as Sesshuu (1420 (Oei Year 27) - 1506 (Eisho year 3), he was an ink painter and Zen monk active in the Muromachi period in the 2nd half of the 15th
<cmt></cmt>
<e type="trans" ver="2">Known as Sesshuu (1420 - 1506, he was an ink painter and Zen monk active in the Muromachi period in the latter half of the 15th century, and was called a
<cmt>西洋暦のみとする。自然な表現に修正。</cmt>
<e type="check" ver="1">Known as Sesshu (1420 - 1506), he was an ink painter and Zen monk active in the Muromachi period in the latter half of the 15th century, and was called a
<cmt>フィードバックに基づき翻訳を修正しました。</cmt>
</sen>
<sen id="2">
<j>日本の水墨画を一変させた。</j>
<e type="trans" ver="1">He changed Japanese ink painting.</e>
<cmt></cmt>
<e type="trans" ver="2">He changed Japanese ink painting.</e>
<cmt>修正なし</cmt>
<e type="check" ver="1">He revolutionized the Japanese ink painting.</e>
<cmt>フィードバックに基づき翻訳を修正しました。</cmt>
```



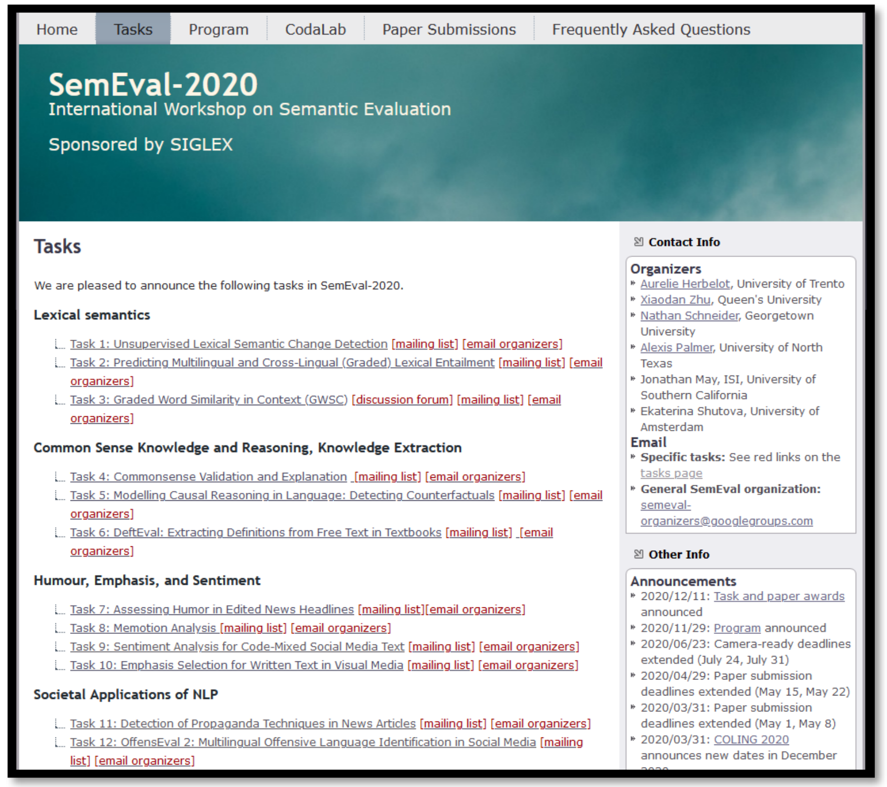
author	date	headlines	read_more	text	ctext
Chhavi Tyagi	03 Aug 2017,Thursday	Daman & Diu revokes mandatory Rakshabandhan in offices order	http://www.hindustantimes.com/india-news/rakshabandhan-compulsory-in-daman-and-diu-women-employees-t...	The Administration of Union Territory Daman and Diu has revoked its order that made it compulsory fo...	The Daman and Diu administration on Wednesday withdrew a circular that asked women staff to tie rakh...
Daisy Mowke	03 Aug 2017,Thursday	Malaika slams user who trolled her for 'divorcing rich man'	http://www.hindustantimes.com/bollywood/malaika-arora-khan-was-trolled-for-divorcing-a-rich-man-her-...	Malaika Arora slammed an Instagram user who trolled her for "divorcing a rich man" and "having fun w...	From her special numbers to TV?appearances, Bollywood actor Malaika Arora Khan has managed to carve ...
Arshiya Chopra	03 Aug 2017,Thursday	'Virgin' now corrected to 'Unmarried' in IGIMS' form	http://www.hindustantimes.com/patna/bihar-igims-form-loses-virginity-after-row-opts-for-unmarried-in...	The Indira Gandhi Institute of Medical Sciences (IGIMS) in Patna on Thursday made corrections in its...	The Indira Gandhi Institute of Medical Sciences (IGIMS) in Patna amended its marital declaration for...
Sumedha Sehra	03 Aug 2017,Thursday	Aaj aapne pakad liya: LeT man Dujana before being killed	http://indiatoday.intoday.in/story/abu-dujana-last-phone-call-lashkar-e-taiba-jammu-and-kashmir/1/10...	Lashkar-e-Taiba's Kashmir commander Abu Dujana, who was killed by security forces, said "Kabhi hum a...	Lashkar-e-Taiba's Kashmir commander Abu Dujana was killed in an encounter in a village in Pulwama di...



type	text
ham	Hope you are having a good week. Just checking in
ham	K..give back my thanks.
ham	Am also doing in cbe only. But have to pay.
spam	complimentary 4 STAR Ibiza Holiday or £10,000 cash needs your URGENT collection. 09066364349 NOW fro...
spam	okmail: Dear Dave this is your final notice to collect your 4* Tenerife Holiday or #5000 CASH award!...

Sample text corpora

- Where to find text corpora?
  - NLP competitions
    - NLP shared task
  - Kaggle
  - Active NLP groups websites



SemEval-2020
International Workshop on Semantic Evaluation

Sponsored by SIGLEX

Tasks

We are pleased to announce the following tasks in SemEval-2020.

Lexical semantics

- Task 1: Unsupervised Lexical Semantic Change Detection [mailing list] [email organizers]
- Task 2: Predicting Multilingual and Cross-Lingual (Graded) Lexical Entailment [mailing list] [email organizers]
- Task 3: Graded Word Similarity in Context (GWSC) [discussion forum] [mailing list] [email organizers]

Common Sense Knowledge and Reasoning, Knowledge Extraction

- Task 4: Commonsense Validation and Explanation [mailing list] [email organizers]
- Task 5: Modelling Causal Reasoning in Language: Detecting Counterfactuals [mailing list] [email organizers]
- Task 6: DeftEval: Extracting Definitions from Free Text in Textbooks [mailing list] [email organizers]

Humour, Emphasis, and Sentiment

- Task 7: Assessing Humor in Edited News Headlines [mailing list][email organizers]
- Task 8: Memotion Analysis [mailing list] [email organizers]
- Task 9: Sentiment Analysis for Code-Mixed Social Media Text [mailing list] [email organizers]
- Task 10: Emphasis Selection for Written Text in Visual Media [mailing list] [email organizers]

Societal Applications of NLP

- Task 11: Detection of Propaganda Techniques in News Articles [mailing list] [email organizers]
- Task 12: OffensEval 2: Multilingual Offensive Language Identification in Social Media [mailing list] [email organizers]

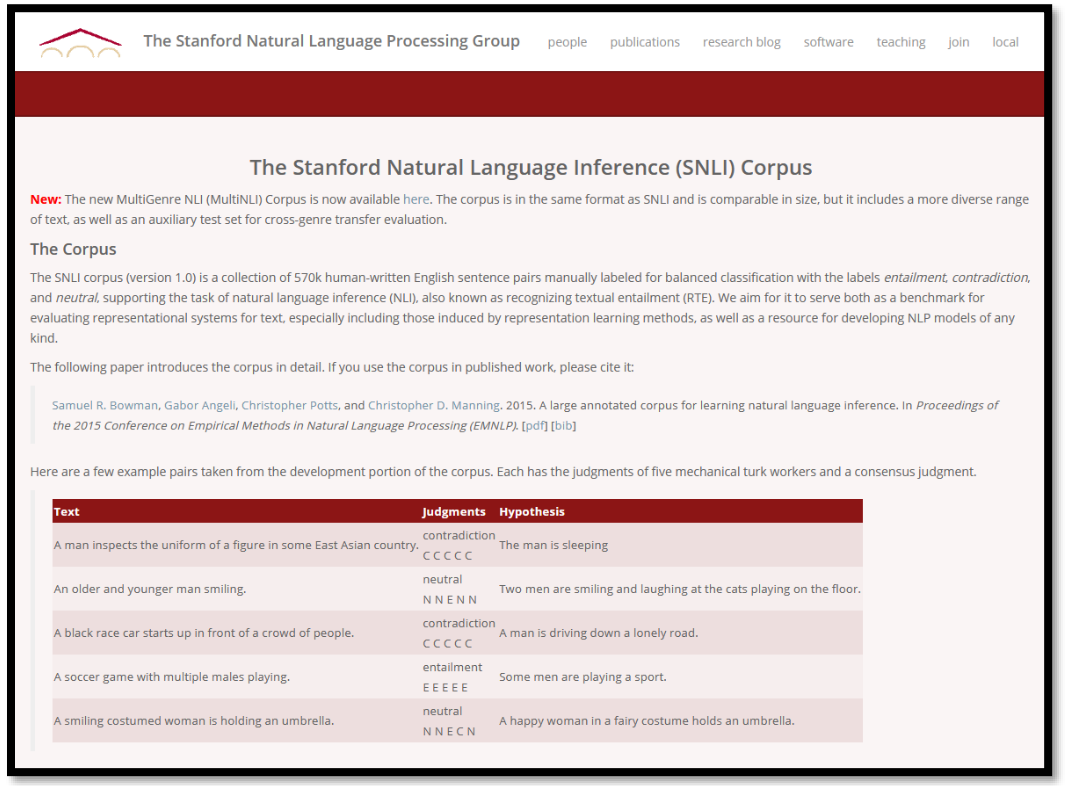
Contact Info

Organizers
- Aurelie Herbelot, University of Trento
- Xiaodan Zhu, Queen's University
- Nathan Schneider, Georgetown University
- Alexis Palmer, University of North Texas
- Jonathan May, ISI, University of Southern California
- Ekaterina Shutova, University of Amsterdam

Email
- **Specific tasks:** See red links on the tasks page
- **General SemEval organization:** semeval-organizers@googlegroups.com

Other Info

Announcements
- 2020/12/11: Task and paper awards announced
- 2020/11/29: Program announced
- 2020/06/23: Camera-ready deadlines extended (July 24, July 31)
- 2020/04/29: Paper submission deadlines extended (May 15, May 22)
- 2020/03/31: Paper submission deadlines extended (May 1, May 8)
- 2020/03/31: COLING 2020 announces new dates in December



The Stanford Natural Language Inference (SNLI) Corpus

New: The new MultiGenre NLI (MultiNLI) Corpus is now available here. The corpus is in the same format as SNLI and is comparable in size, but it includes a more diverse range of text, as well as an auxiliary test set for cross-genre transfer evaluation.

The Corpus

The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels *entailment, contradiction*, and *neutral*, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE). We aim for it to serve both as a benchmark for evaluating representational systems for text, especially including those induced by representation learning methods, as well as a resource for developing NLP models of any kind.

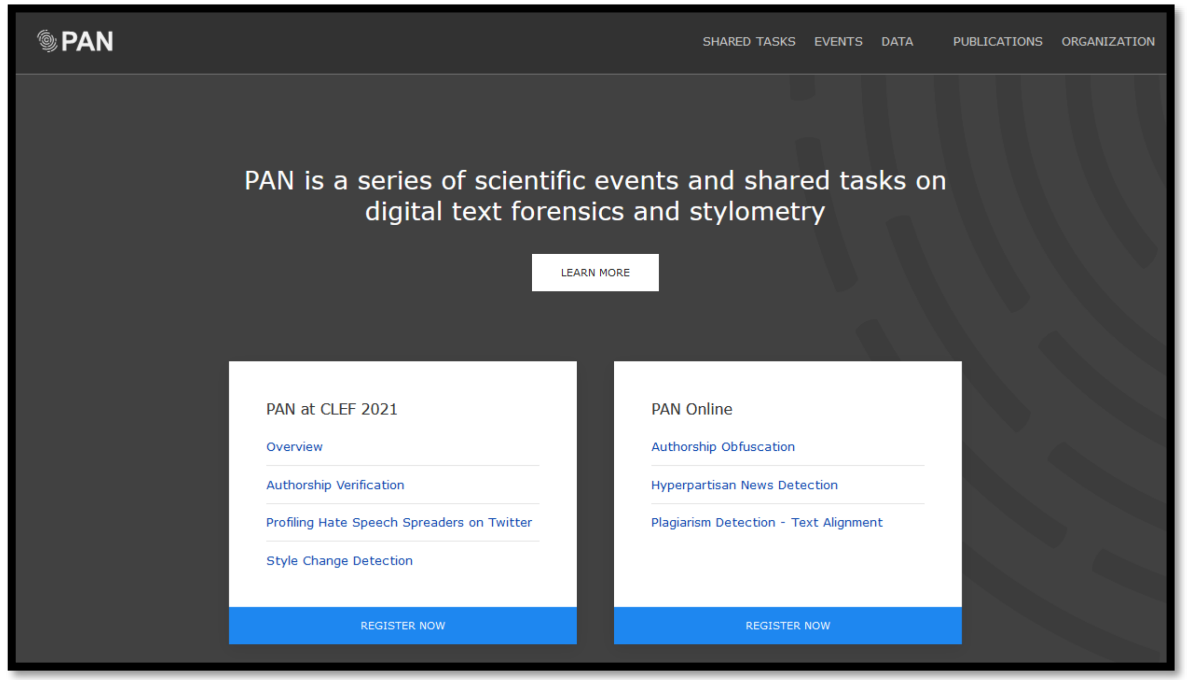
The following paper introduces the corpus in detail. If you use the corpus in published work, please cite it:

> Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).* [pdf] [bib]

Here are a few example pairs taken from the development portion of the corpus. Each has the judgments of five mechanical turk workers and a consensus judgment.

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

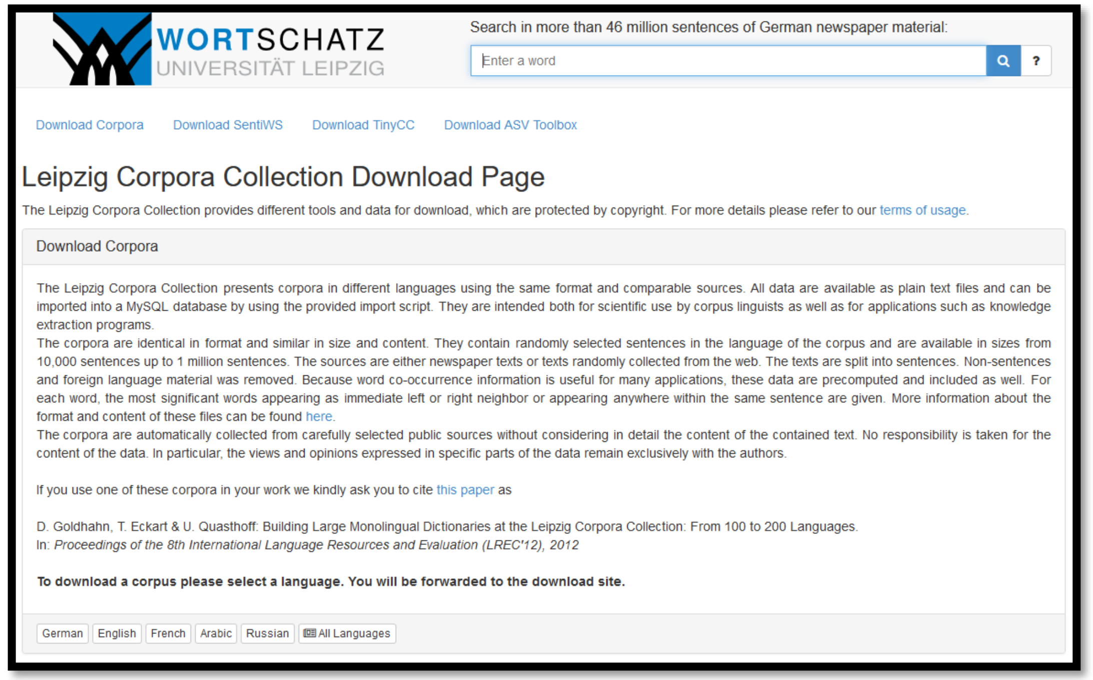
kaggle



PAN is a series of scientific events and shared tasks on digital text forensics and stylometry

PAN at CLEF 2021
- Overview
- Authorship Verification
- Profiling Hate Speech Spreaders on Twitter
- Style Change Detection

PAN Online
- Authorship Obfuscation
- Hyperpartisan News Detection
- Plagiarism Detection - Text Alignment



Search in more than 46 million sentences of German newspaper material:

Leipzig Corpora Collection Download Page

The Leipzig Corpora Collection provides different tools and data for download, which are protected by copyright. For more details please refer to our terms of usage.

Download Corpora

The Leipzig Corpora Collection presents corpora in different languages using the same format and comparable sources. All data are available as plain text files and can be imported into a MySQL database by using the provided import script. They are intended both for scientific use by corpus linguists as well as for applications such as knowledge extraction programs.
The corpora are identical in format and similar in size and content. They contain randomly selected sentences in the language of the corpus and are available in sizes from 10,000 sentences up to 1 million sentences. The sources are either newspaper texts or texts randomly collected from the web. The texts are split into sentences. Non-sentences and foreign language material was removed. Because word co-occurrence information is useful for many applications, these data are precomputed and included as well. For each word, the most significant words appearing as immediate left or right neighbor or appearing anywhere within the same sentence are given. More information about the format and content of these files can be found here.
The corpora are automatically collected from carefully selected public sources without considering in detail the content of the contained text. No responsibility is taken for the content of the data. In particular, the views and opinions expressed in specific parts of the data remain exclusively with the authors.

If you use one of these corpora in your work we kindly ask you to cite this paper as

D. Goldhahn, T. Eckart & U. Quasthoff: Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages.
In: *Proceedings of the 8th International Language Resources and Evaluation (LREC'12), 2012*

To download a corpus please select a language. You will be forwarded to the download site.

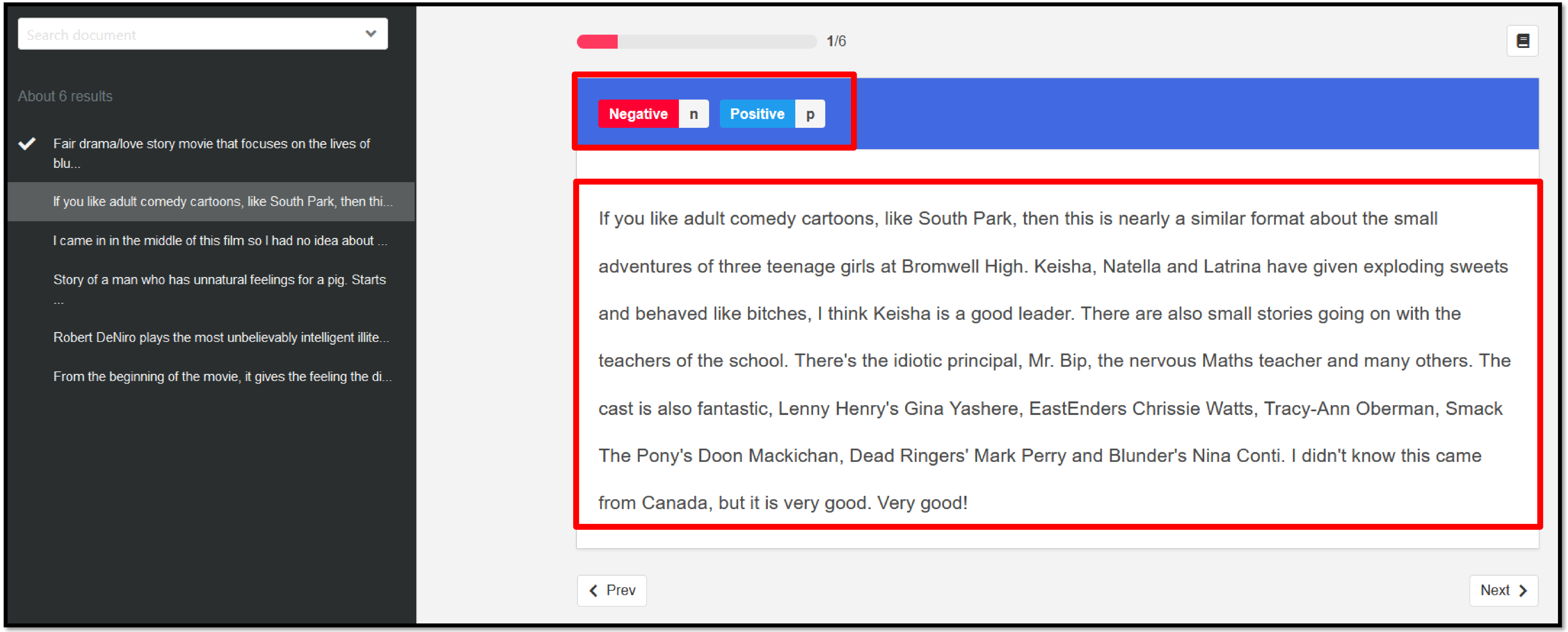
Text corpus

- What is a text corpus?
- Sample text corpora
- Corpus annotation
- Underfitting and overfitting
- Data splitting
- Text corpora in Python

Corpus annotation

- It would happen that you have to compile your own corpus for the task and domain of interests
  - e.g., generating a text summarization corpus for medical texts in German
- Apart from the pure text, a corpus can also be provided with additional linguistic information, called **annotation**
- Corpus annotation is the practice of adding interpretative linguistic information to a piece of text

Corpus annotation

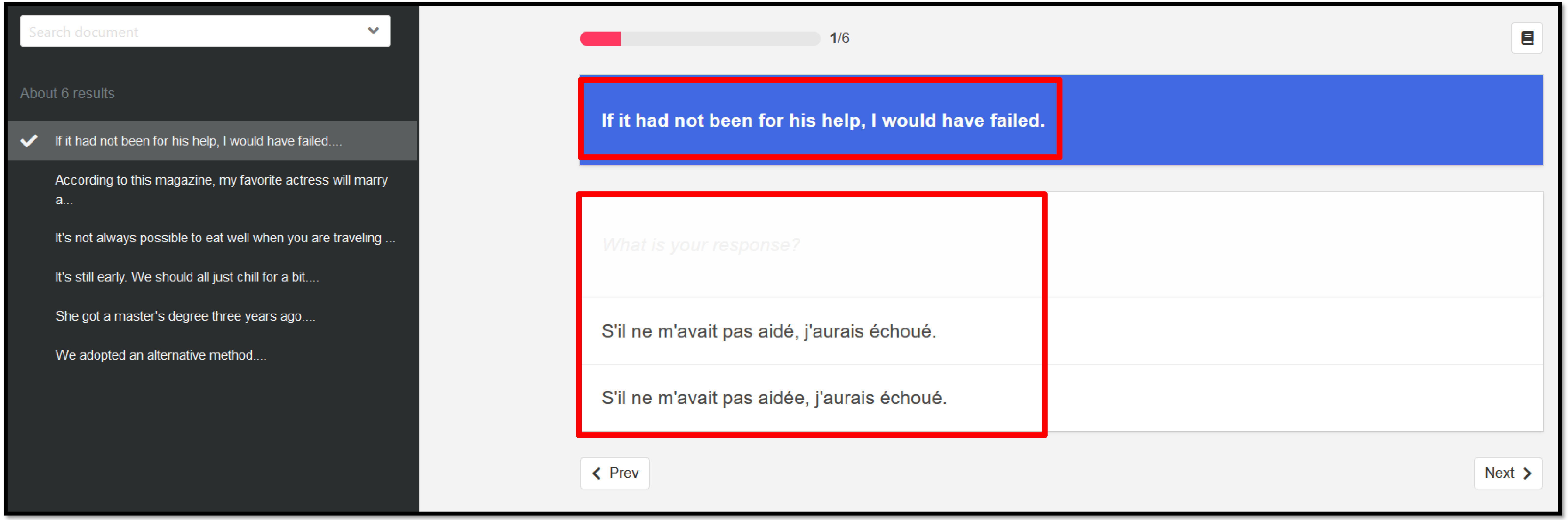


About 6 results

- ✓ Fair drama/love story movie that focuses on the lives of blu...
- If you like adult comedy cartoons, like South Park, then thi...
- I came in in the middle of this film so I had no idea about ...
- Story of a man who has unnatural feelings for a pig. Starts ...
- Robert DeNiro plays the most unbelievably intelligent illite...
- From the beginning of the movie, it gives the feeling the di...

If you like adult comedy cartoons, like South Park, then this is nearly a similar format about the small adventures of three teenage girls at Bromwell High. Keisha, Natella and Latrina have given exploding sweets and behaved like bitches, I think Keisha is a good leader. There are also small stories going on with the teachers of the school. There's the idiotic principal, Mr. Bip, the nervous Maths teacher and many others. The cast is also fantastic, Lenny Henry's Gina Yashere, EastEnders Chrissie Watts, Tracy-Ann Oberman, Smack The Pony's Doon Mackichan, Dead Ringers' Mark Perry and Blunder's Nina Conti. I didn't know this came from Canada, but it is very good. Very good!

Corpus annotation



✓ If it had not been for his help, I would have failed....

According to this magazine, my favorite actress will marry a...

It's not always possible to eat well when you are traveling ...

It's still early. We should all just chill for a bit....

She got a master's degree three years ago....

We adopted an alternative method....

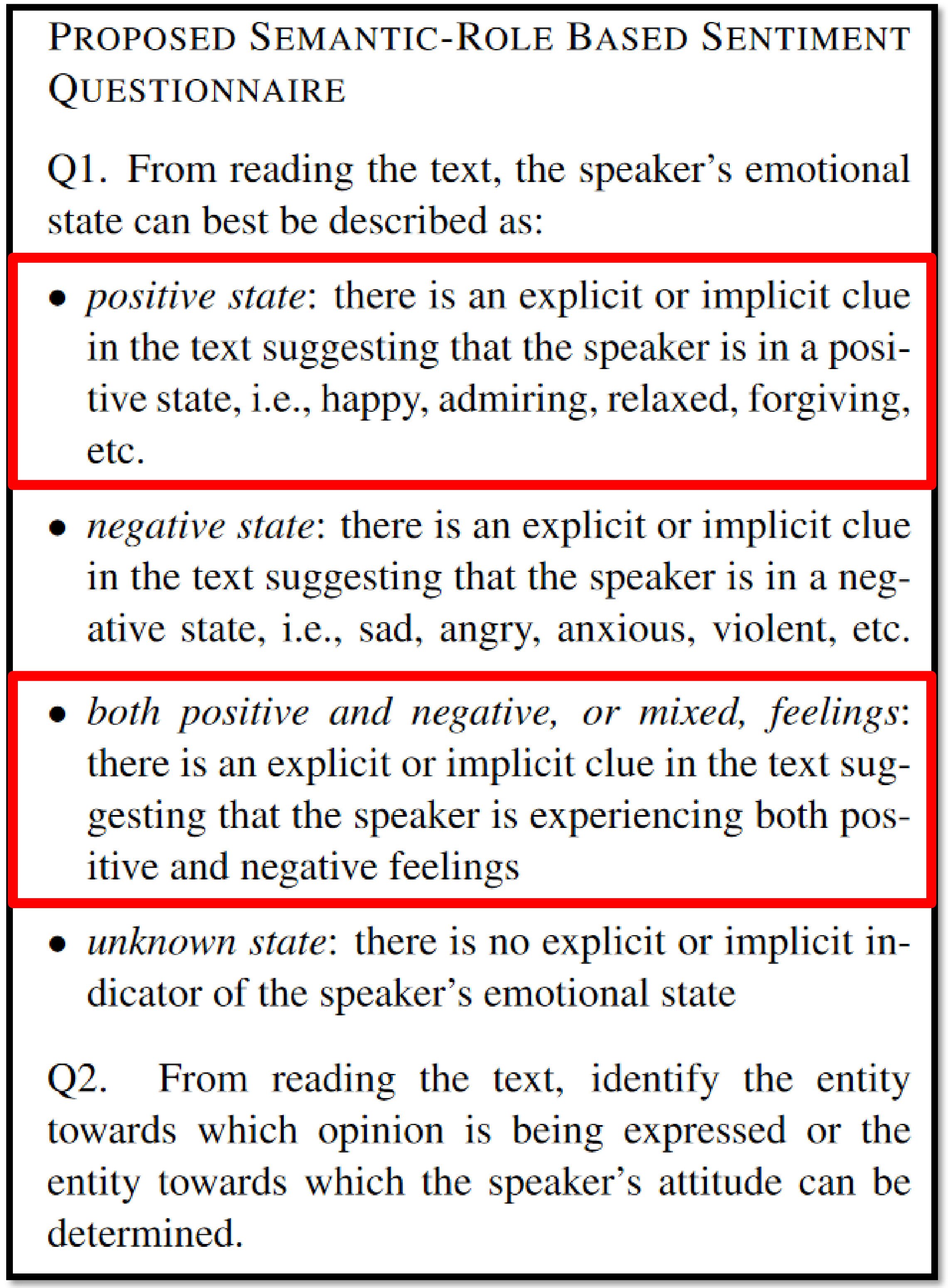
If it had not been for his help, I would have failed.

S'il ne m'avait pas aidé, j'aurais échoué.

S'il ne m'avait pas aidée, j'aurais échoué.

Corpus annotation

- The annotation process could meet lots of questions and ambiguities
- Annotation guideline
  - Describe the annotation procedure as generic as possible but as precise as necessary
  - So that human annotators can annotate the concept or phenomenon in any text without running into problems or ambiguity issues



PROPOSED SEMANTIC-ROLE BASED SENTIMENT QUESTIONNAIRE

Q1. From reading the text, the speaker's emotional state can best be described as:

- *positive state*: there is an explicit or implicit clue in the text suggesting that the speaker is in a positive state, i.e., happy, admiring, relaxed, forgiving, etc.
- *negative state*: there is an explicit or implicit clue in the text suggesting that the speaker is in a negative state, i.e., sad, angry, anxious, violent, etc.
- *both positive and negative, or mixed, feelings*: there is an explicit or implicit clue in the text suggesting that the speaker is experiencing both positive and negative feelings
- *unknown state*: there is no explicit or implicit indicator of the speaker's emotional state

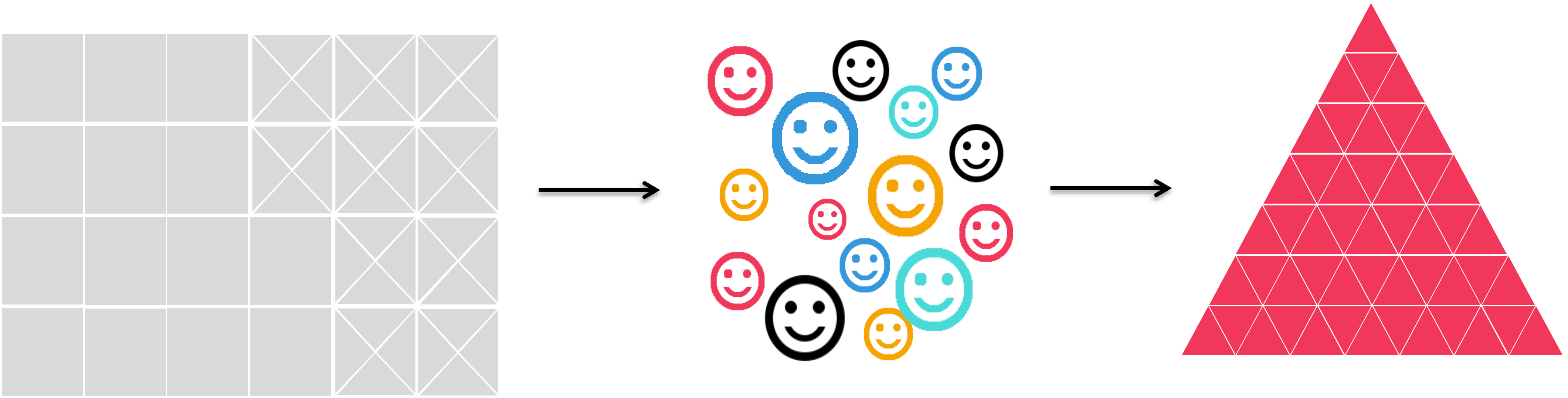
Q2. From reading the text, identify the entity towards which opinion is being expressed or the entity towards which the speaker's attitude can be determined.

Corpus annotation

- Laboratory vs. crowdsourcing
  - Laboratory
    - asking experts (e.g., linguists) for data annotation
  - Crowdsourcing
    - Is a participatory method of building a dataset with the help of a large group of people

Corpus annotation

- Crowdsourcing



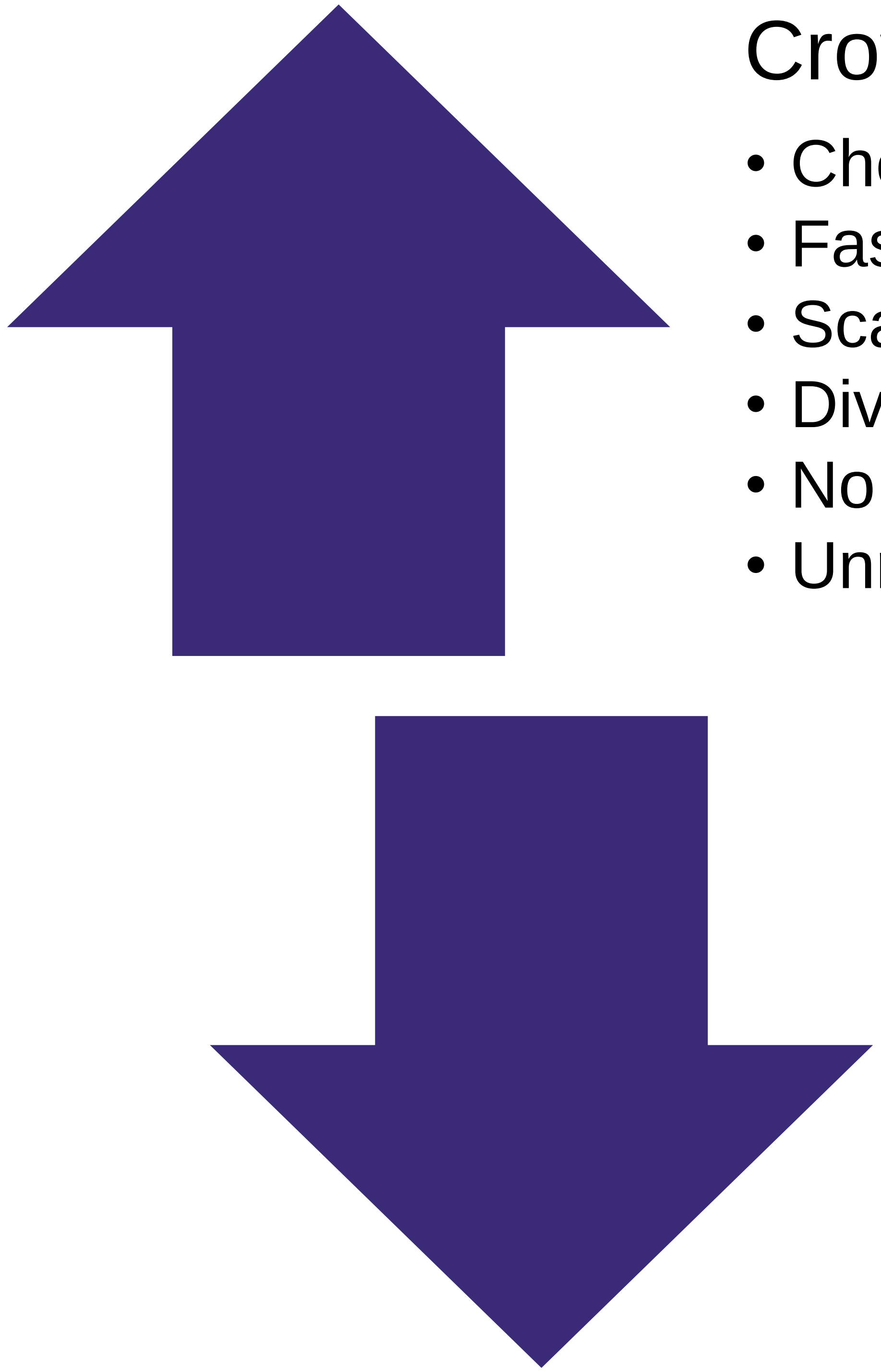
amazon mechanical turk
beta
Artificial Artificial Intelligence

Crowdee

microWorkers
work & earn or offer a micro job

CrowdFlower

Corpus annotation



Crowdsourcing

- Cheaper
- Faster
- Scalable
- Diverse group of participants
- No moderator
- Unreliable data

Laboratory

- High reliability
- Controlled unwanted factors
- Environment constant, no noise
- Participants: one-to-one contact to moderator

Text corpus

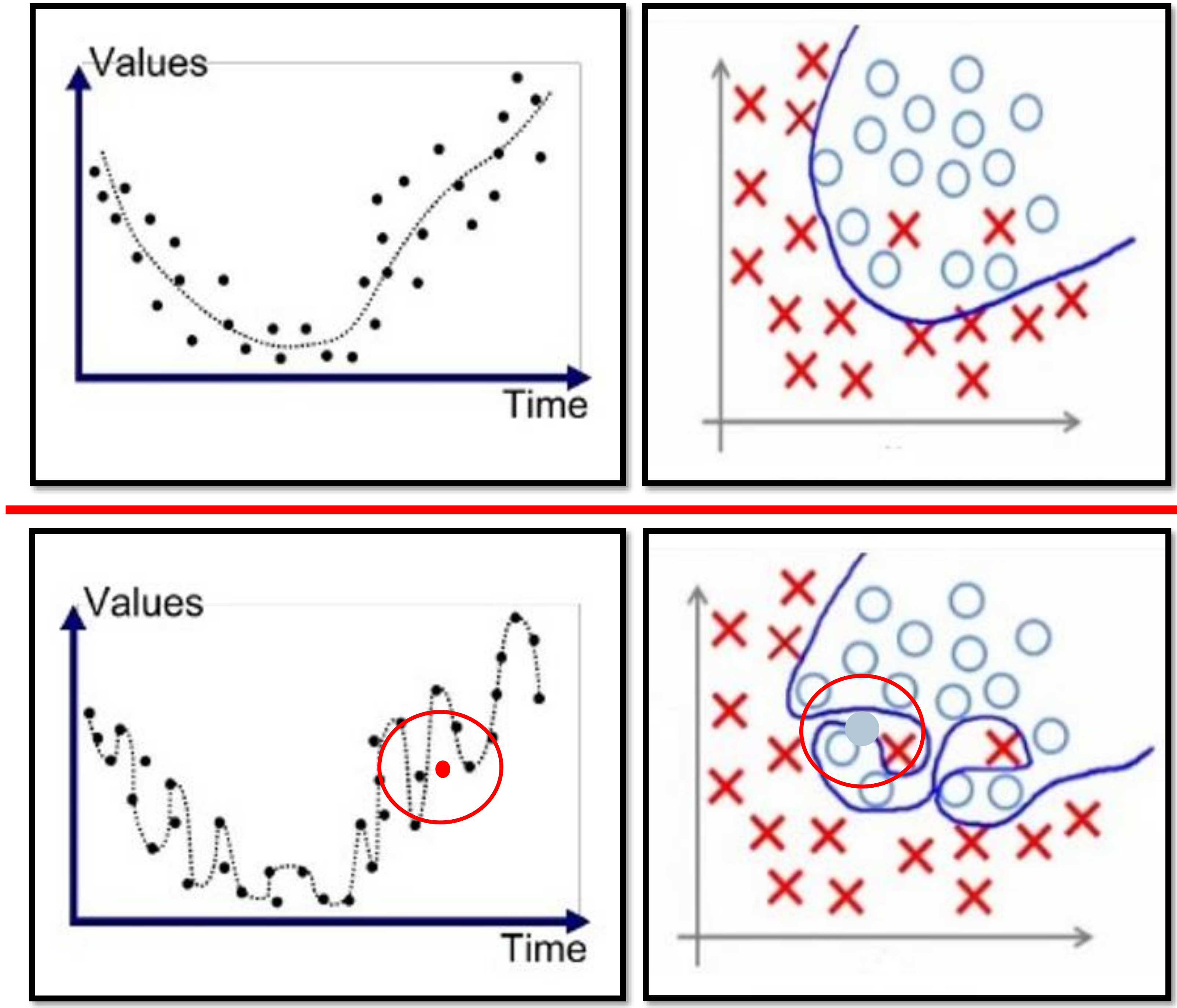
- What is a text corpus?
- Sample text corpora
- Corpus annotation
- Underfitting and overfitting
- Data splitting
- Text corpora in Python

Underfitting and overfitting

- Overfitting
  - Overfitting refers to a model that models the training data too well
  - It happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data
  - This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model
  - The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize

Underfitting and overfitting

- Overfitting



Appropriate-fitting

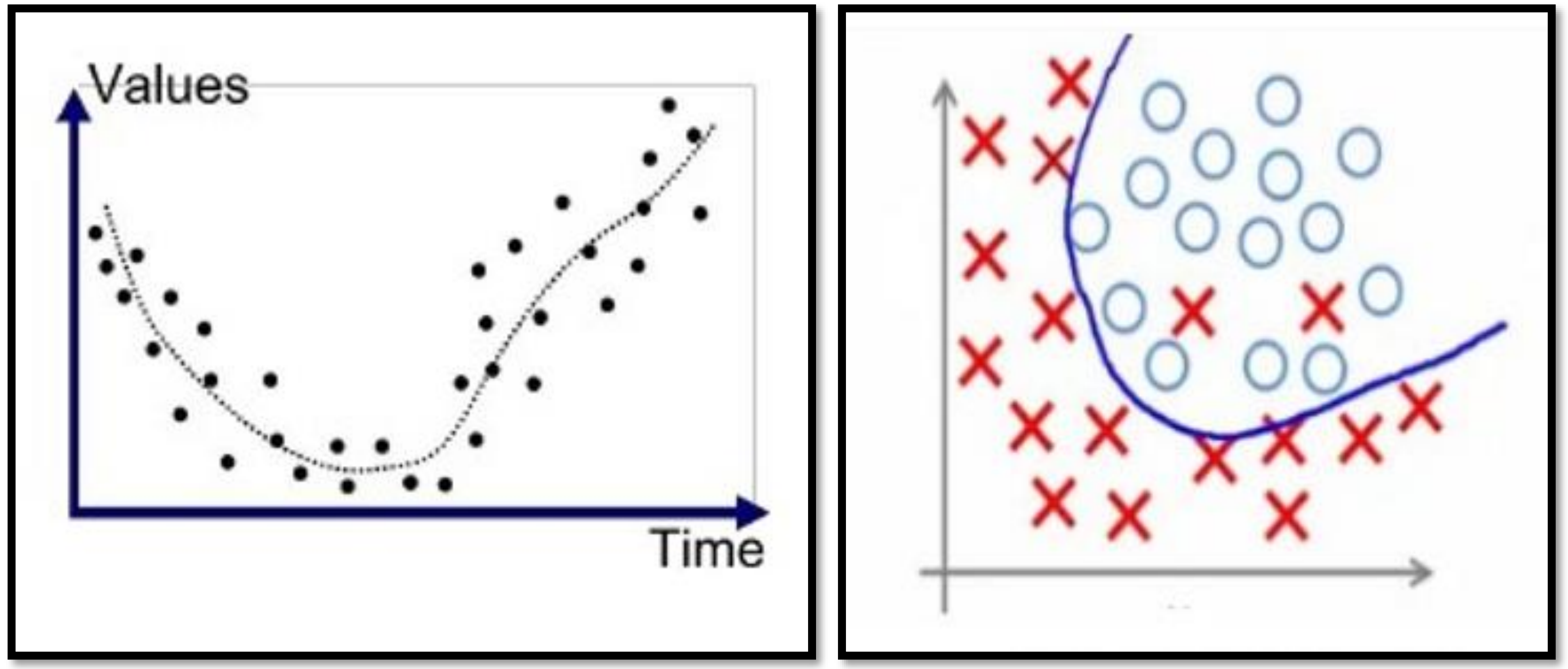
Overfitting

Underfitting and overfitting

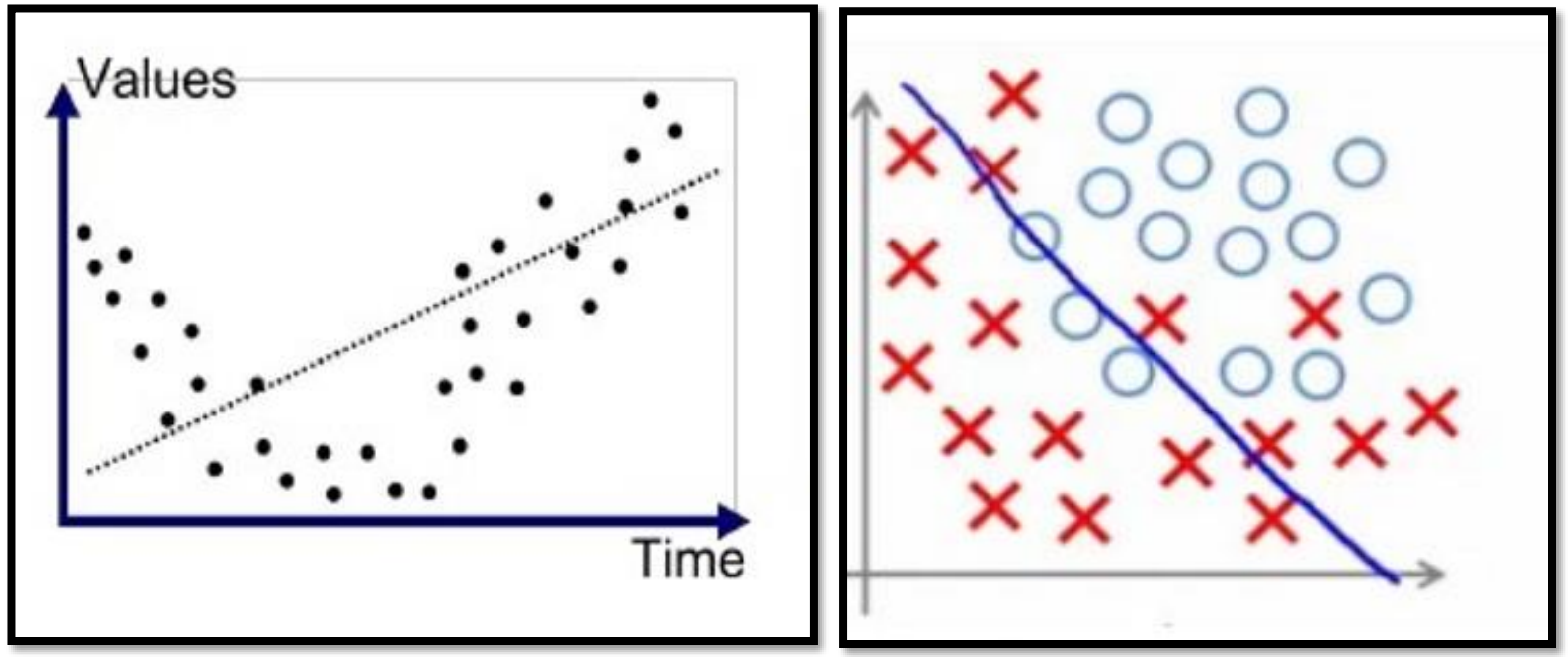
- Underfitting
  - Underfitting refers to a model that can neither model the training data nor generalize to new data
  - An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data

Underfitting and overfitting

- Underfitting



Appropriate-fitting



Underfitting

Underfitting and overfitting

- Overfitting
  - Increase training data
  - Reduce model complexity
  - Cross-validation
  - Early stopping during the training phase
  - Regularization
- Underfitting
  - Increase model complexity
  - Increase number of features
  - Remove noise from the data

Text corpus

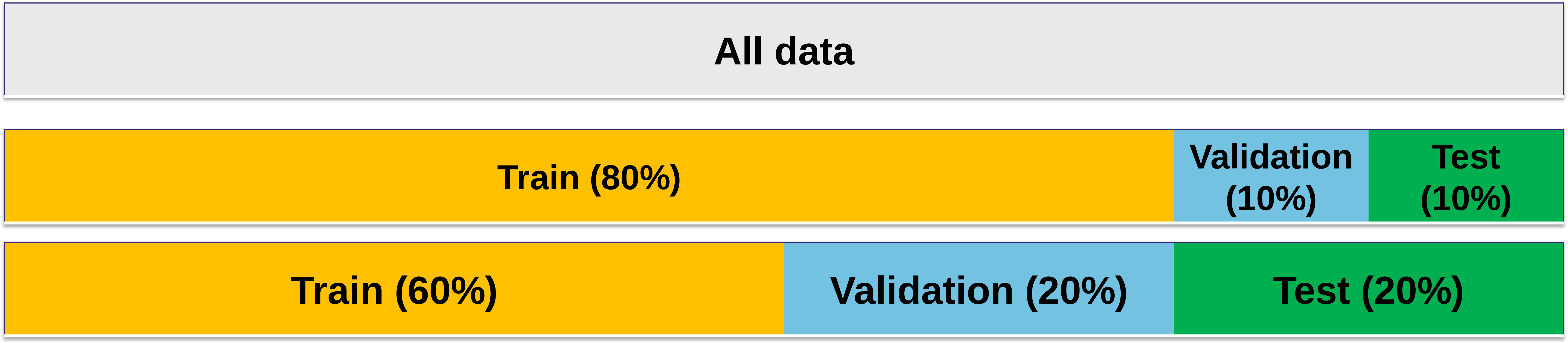
- What is a text corpus?
- Sample text corpora
- Corpus annotation
- Underfitting and overfitting
- Data splitting
- Text corpora in Python

Data splitting

- As a common practice in machine learning (NLP), we have to split the dataset (corpus) into different part in order to train a model and test it
- Two main approaches for splitting data
  - Train / Validation / Test
  - Cross validation

Data splitting

- Train / Validation / Test
  - **Training Dataset**: The sample of data used to fit the model
  - **Validation Dataset**: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters
  - **Test Dataset**: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset



All data

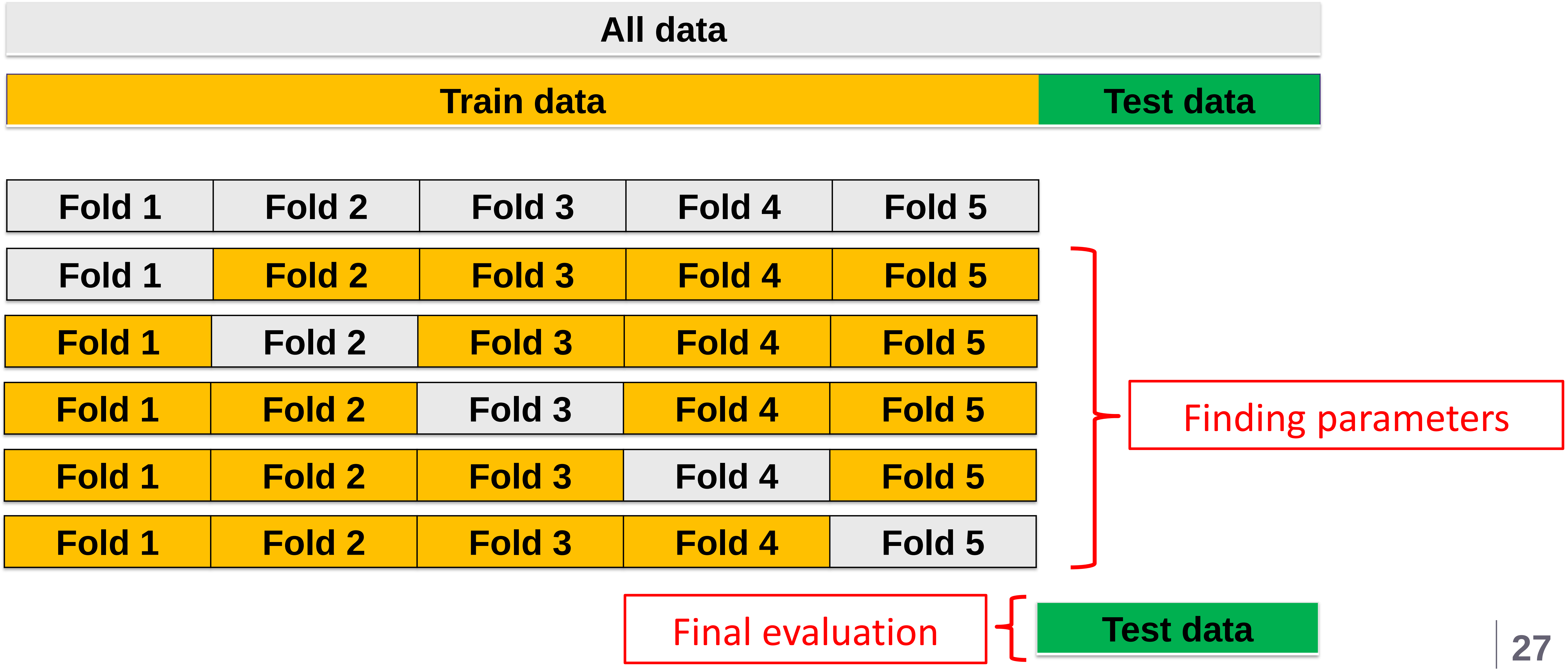
Train (80%) | Validation (10%) | Test (10%)

Train (60%) | Validation (20%) | Test (20%)

Data splitting

- Cross validation
  - Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample
- The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into
  - the procedure is often called k-fold cross-validation

Data splitting



All data

Train data | Test data

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Finding parameters

Final evaluation

Test data

Text corpus

- What is a text corpus?
- Sample text corpora
- Corpus annotation
- Underfitting and overfitting
- Data splitting
- Text corpora in Python

Text corpora in Python

- Reuters Corpus
- Brown Corpus
- Web and Chat Text
- Gutenberg Corpus
- …

- And Corpora in Other Languages

NLTK
Natural Language Toolkit

```
>>> from nltk.corpus import reuters
>>> reuters.categories()
['acq', 'alum', 'barley', 'bop',
'carcass', 'castor-oil', 'cocoa',
'coconut', 'coconut-oil', 'coffee’, ...]
>>> from nltk.corpus import brown
>>> brown.words()
['The', 'Fulton', 'County', 'Grand',
'Jury', 'said', ...]
```

Summary

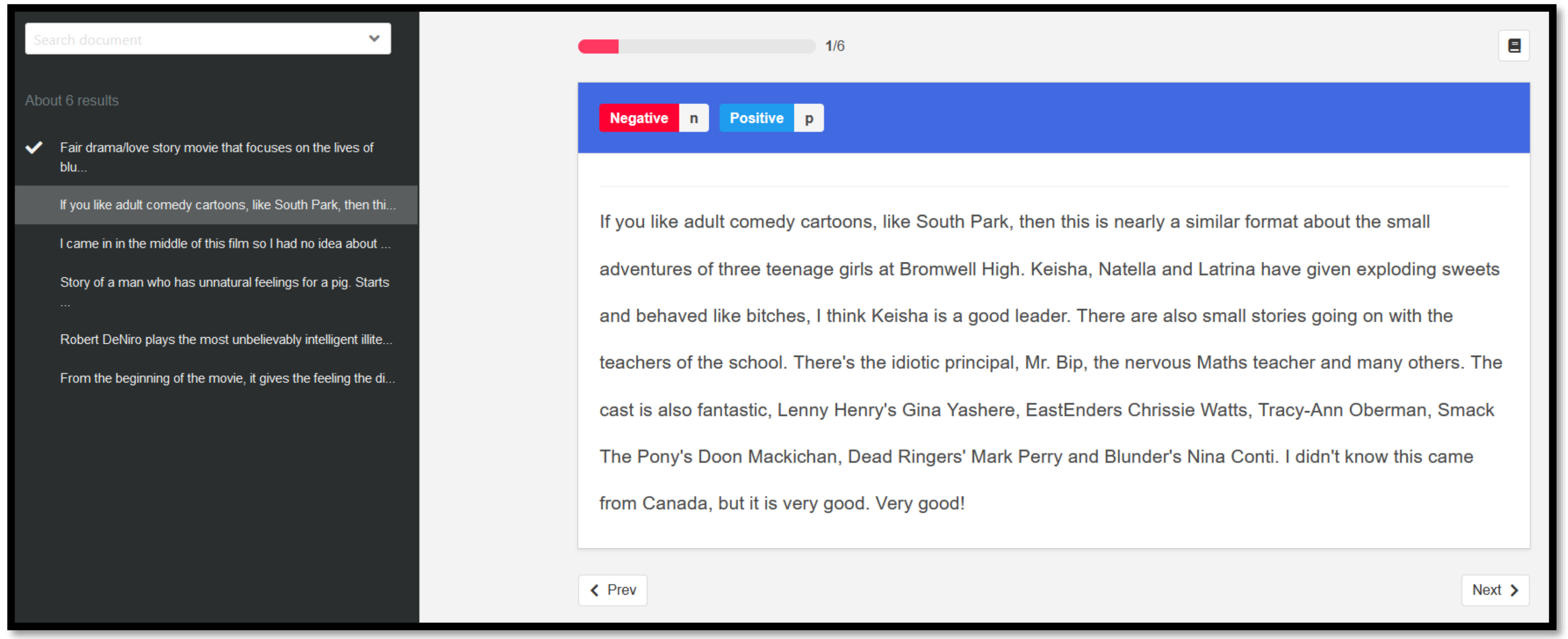
- A corpus is a collection of texts, **written** or spoken

amazonmechanicalturk beta
Artificial Artificial Intelligence

Crowdee

microWorkers
work & earn or offer a micro job

CrowdFlower



✓ Fair drama/love story movie that focuses on the lives of blu...

If you like adult comedy cartoons, like South Park, then thi...

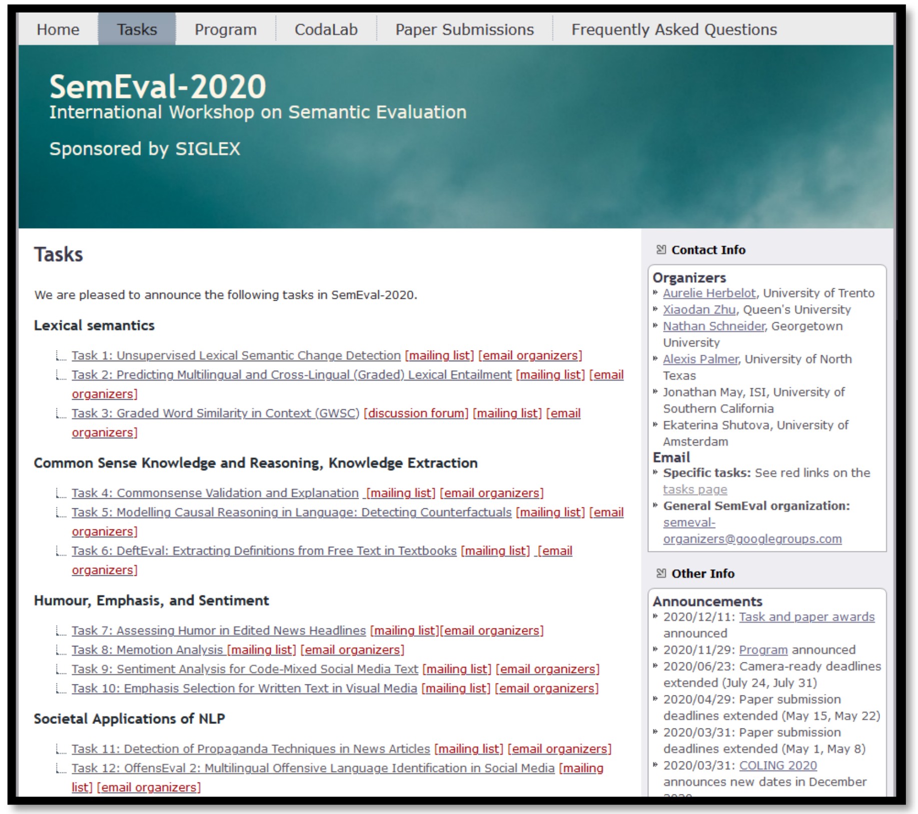
I came in in the middle of this film so I had no idea about ...

Story of a man who has unnatural feelings for a pig. Starts ...

Robert DeNiro plays the most unbelievably intelligent illite...

From the beginning of the movie, it gives the feeling the di...

If you like adult comedy cartoons, like South Park, then this is nearly a similar format about the small adventures of three teenage girls at Bromwell High. Keisha, Natella and Latrina have given exploding sweets and behaved like bitches, I think Keisha is a good leader. There are also small stories going on with the teachers of the school. There's the idiotic principal, Mr. Bip, the nervous Maths teacher and many others. The cast is also fantastic, Lenny Henry's Gina Yashere, EastEnders Chrissie Watts, Tracy-Ann Oberman, Smack The Pony's Doon Mackichan, Dead Ringers' Mark Perry and Blunder's Nina Conti. I didn't know this came from Canada, but it is very good. Very good!



SemEval-2020
International Workshop on Semantic Evaluation

Sponsored by SIGLEX

Tasks

We are pleased to announce the following tasks in SemEval-2020.

Lexical semantics

- Task 1: Unsupervised Lexical Semantic Change Detection [mailing list] [email organizers]
- Task 2: Predicting Multilingual and Cross-Lingual (Graded) Lexical Entailment [mailing list] [email organizers]
- Task 3: Graded Word Similarity in Context (GWSC) [discussion forum] [mailing list] [email organizers]

Common Sense Knowledge and Reasoning, Knowledge Extraction

- Task 4: Commonsense Validation and Explanation [mailing list] [email organizers]
- Task 5: Modelling Causal Reasoning in Language: Detecting Counterfactuals [mailing list] [email organizers]
- Task 6: DeftEval: Extracting Definitions from Free Text in Textbooks [mailing list] [email organizers]

Humour, Emphasis, and Sentiment

- Task 7: Assessing Humor in Edited News Headlines [mailing list][email organizers]
- Task 8: Memotion Analysis [mailing list] [email organizers]
- Task 9: Sentiment Analysis for Code-Mixed Social Media Text [mailing list] [email organizers]
- Task 10: Emphasis Selection for Written Text in Visual Media [mailing list] [email organizers]

Societal Applications of NLP

- Task 11: Detection of Propaganda Techniques in News Articles [mailing list] [email organizers]
- Task 12: OffensEval 2: Multilingual Offensive Language Identification in Social Media [mailing list] [email organizers]

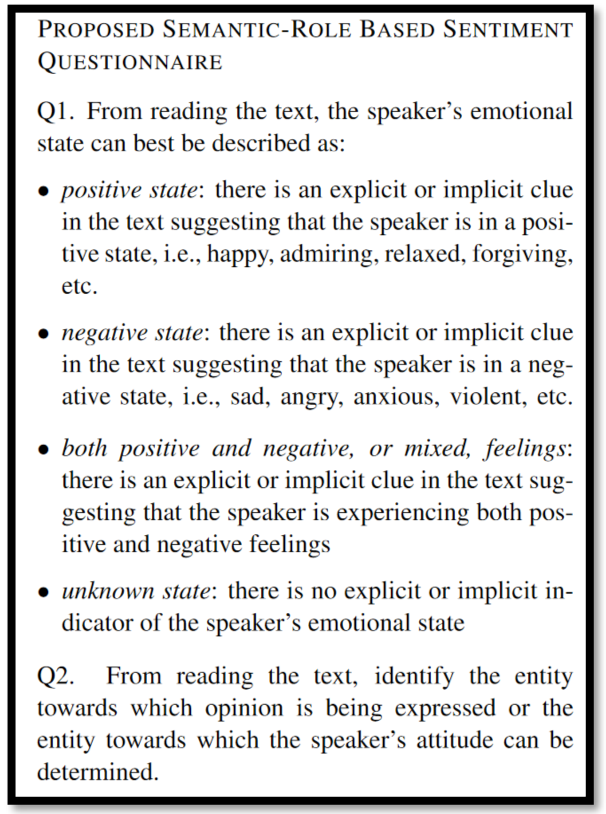
Contact Info

Organizers
- Aurelie Herbelot, University of Trento
- Xiaodan Zhu, Queen's University
- Nathan Schneider, Georgetown University
- Alexis Palmer, University of North Texas
- Jonathan May, ISI, University of Southern California
- Ekaterina Shutova, University of Amsterdam

Email
- **Specific tasks:** See red links on the tasks page
- **General SemEval organization:** semeval-organizers@googlegroups.com

Other Info

Announcements
- 2020/12/11: Task and paper awards announced
- 2020/11/29: Program announced
- 2020/06/23: Camera-ready deadlines extended (July 24, July 31)
- 2020/04/29: Paper submission deadlines extended (May 15, May 22)
- 2020/03/31: Paper submission deadlines extended (May 1, May 8)
- 2020/03/31: COLING 2020 announces new dates in December

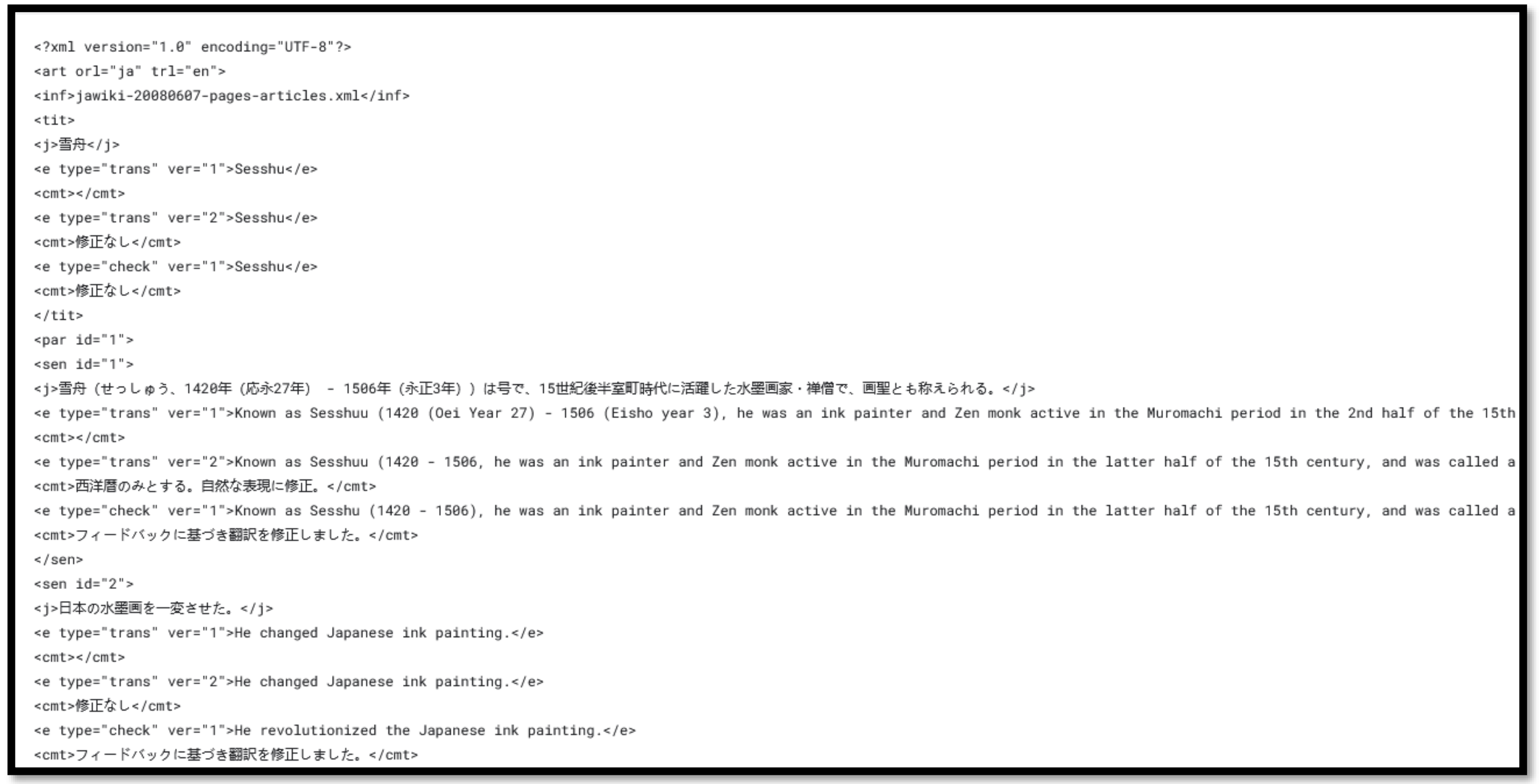


PROPOSED SEMANTIC-ROLE BASED SENTIMENT QUESTIONNAIRE

Q1. From reading the text, the speaker's emotional state can best be described as:

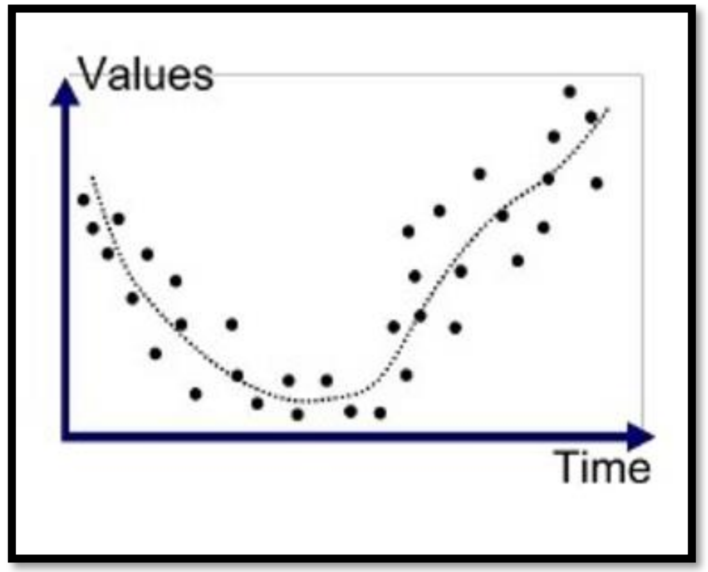
- *positive state*: there is an explicit or implicit clue in the text suggesting that the speaker is in a positive state, i.e., happy, admiring, relaxed, forgiving, etc.
- *negative state*: there is an explicit or implicit clue in the text suggesting that the speaker is in a negative state, i.e., sad, angry, anxious, violent, etc.
- *both positive and negative, or mixed, feelings*: there is an explicit or implicit clue in the text suggesting that the speaker is experiencing both positive and negative feelings
- *unknown state*: there is no explicit or implicit indicator of the speaker's emotional state

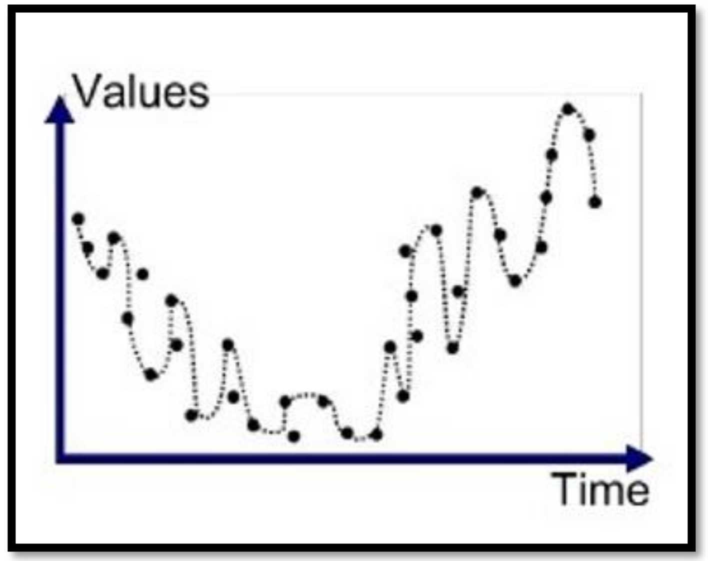
Q2. From reading the text, identify the entity towards which opinion is being expressed or the entity towards which the speaker's attitude can be determined.

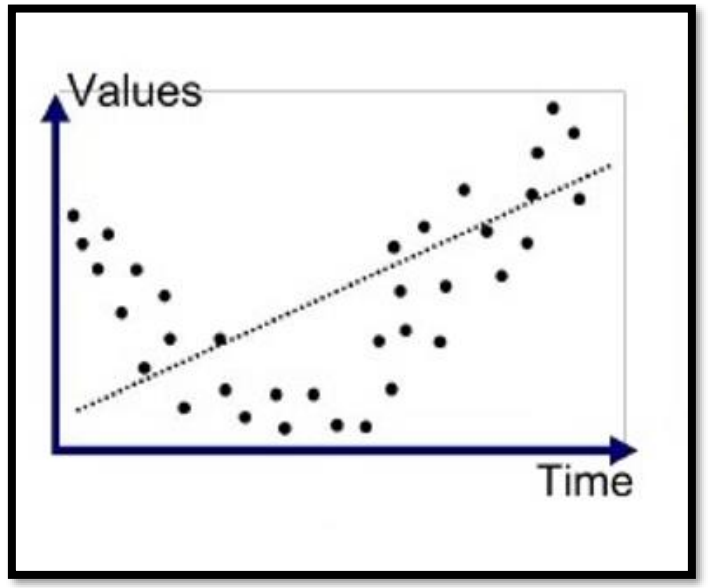


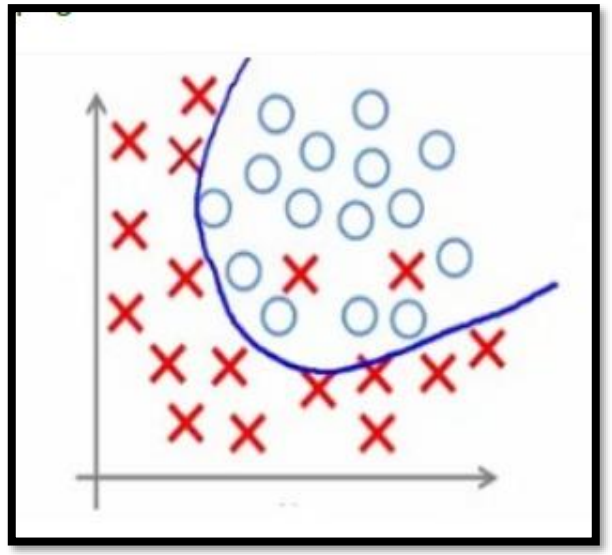
```
<?xml version="1.0" encoding="UTF-8"?>
<art orl="ja" trl="en">
<inf>jawiki-20080607-pages-articles.xml</inf>
<tit>
<j>雪舟</j>
<e type="trans" ver="1">Sesshu</e>
<cmt></cmt>
<e type="trans" ver="2">Sesshu</e>
<cmt>修正なし</cmt>
<e type="check" ver="1">Sesshu</e>
<cmt>修正なし</cmt>
</tit>
<par id="1">
<sen id="1">
<j>雪舟（せっしゅう、1420年（応永27年） - 1506年（永正3年））は号で、15世紀後半室町時代に活躍した水墨画家・禅僧で、画聖とも称えられる。</j>
<e type="trans" ver="1">Known as Sesshuu (1420 (Oei Year 27) - 1506 (Eisho year 3), he was an ink painter and Zen monk active in the Muromachi period in the 2nd half of the 15th
<cmt></cmt>
<e type="trans" ver="2">Known as Sesshuu (1420 - 1506, he was an ink painter and Zen monk active in the Muromachi period in the latter half of the 15th century, and was called a
<cmt>西洋暦のみとする。自然な表現に修正。</cmt>
<e type="check" ver="1">Known as Sesshu (1420 - 1506), he was an ink painter and Zen monk active in the Muromachi period in the latter half of the 15th century, and was called a
<cmt>フィードバックに基づき翻訳を修正しました。</cmt>
</sen>
<sen id="2">
<j>日本の水墨画を一変させた。</j>
<e type="trans" ver="1">He changed Japanese ink painting.</e>
<cmt></cmt>
<e type="trans" ver="2">He changed Japanese ink painting.</e>
<cmt>修正なし</cmt>
<e type="check" ver="1">He revolutionized the Japanese ink painting.</e>
<cmt>フィードバックに基づき翻訳を修正しました。</cmt>
```

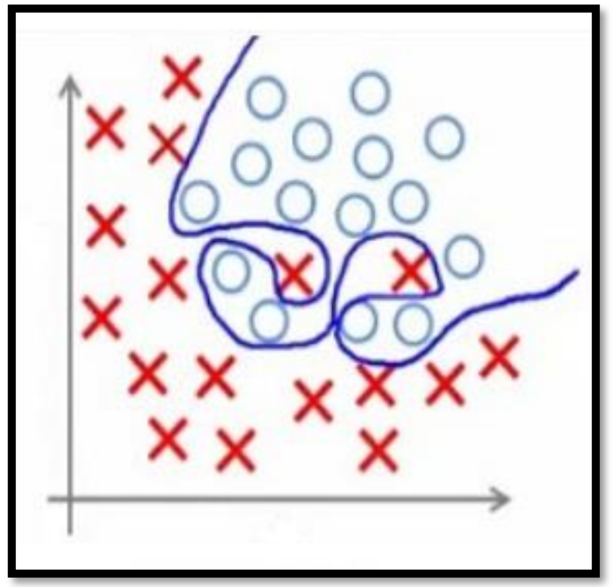
Summary

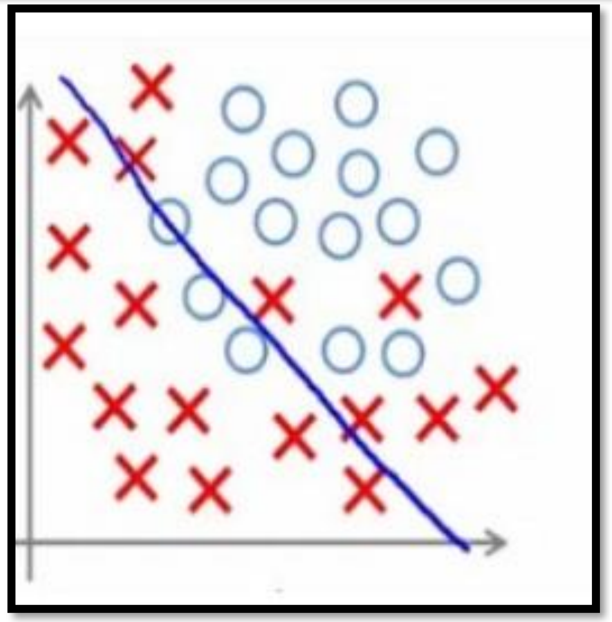




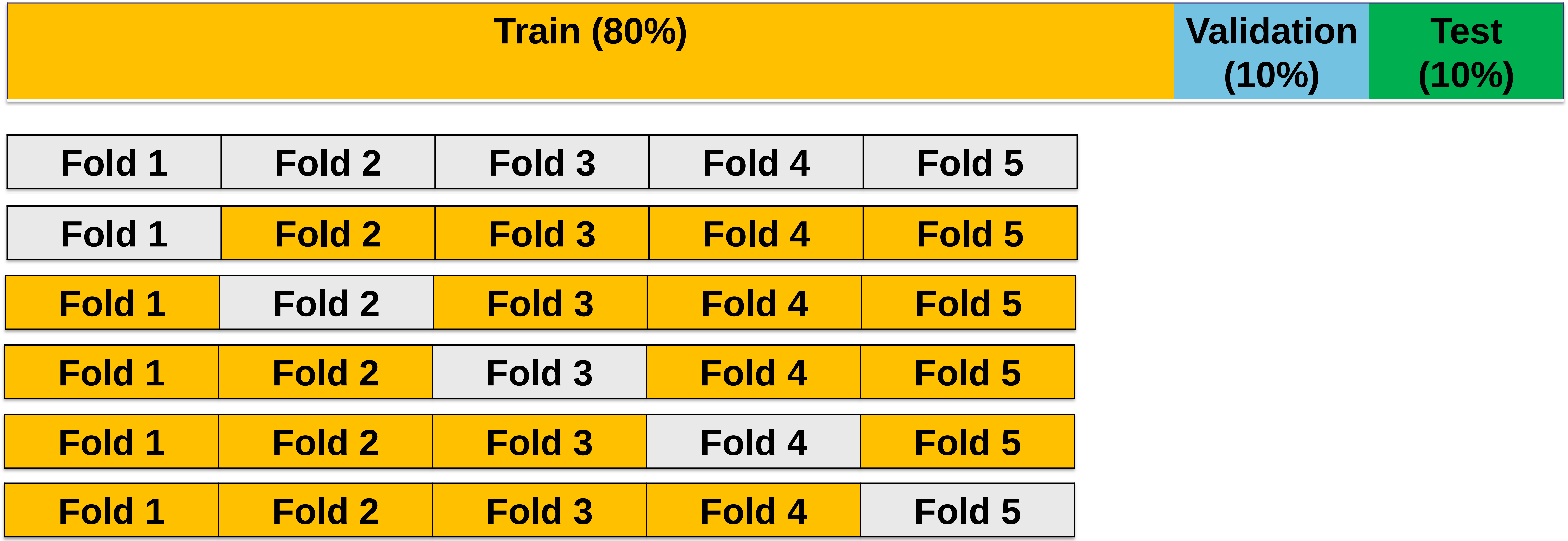








Summary



Train (80%)

Validation (10%)

Test (10%)

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5