# Evaluating content selection in summarization: The pyramid method

Ani Nenkova and Rebecca Passonneau

# Summary

**- Problem**

Summarization techniques are evaluated against golden summaries, carefully built by human annotators but there is no best golden summary

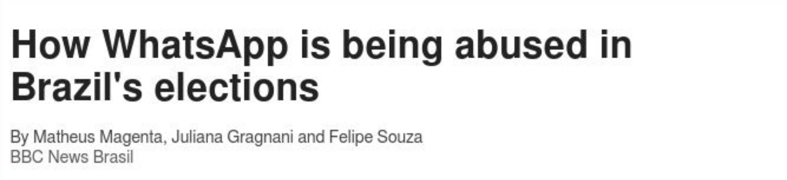
**- Contribution**

Pyramid method: approach that can relatively assess the quality of a given summary against the ones produced by a set of annotators

**- Evaluation**

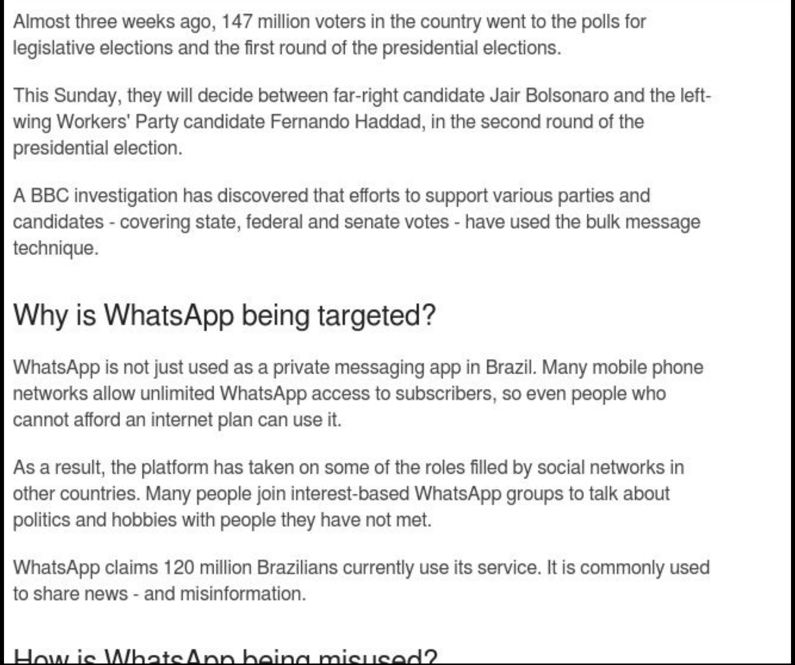
Comparison of scores computed by the pyramid method against scores from the DUC community

# Problem



## How WhatsApp is being abused in Brazil's elections

By Matheus Magenta, Juliana Gragnani and Felipe Souza
BBC News Brasil



Almost three weeks ago, 147 million voters in the country went to the polls for legislative elections and the first round of the presidential elections.

This Sunday, they will decide between far-right candidate Jair Bolsonaro and the left-wing Workers' Party candidate Fernando Haddad, in the second round of the presidential election.

A BBC investigation has discovered that efforts to support various parties and candidates - covering state, federal and senate votes - have used the bulk message technique.

### Why is WhatsApp being targeted?

WhatsApp is not just used as a private messaging app in Brazil. Many mobile phone networks allow unlimited WhatsApp access to subscribers, so even people who cannot afford an internet plan can use it.

As a result, the platform has taken on some of the roles filled by social networks in other countries. Many people join interest-based WhatsApp groups to talk about politics and hobbies with people they have not met.

WhatsApp claims 120 million Brazilians currently use its service. It is commonly used to share news - and misinformation.

### How is WhatsApp being misused?

Political campaigners in Brazil have used software that scrapes Facebook for citizens' phone numbers, and then automatically sends them WhatsApp messages and adds them to WhatsApp groups.

For its part, Facebook says it has banned hundreds of thousands of suspicious WhatsApp accounts believed to be spreading fake news.

*Model summary*

Political campaigners in Brazil have used software that scrapes Facebook for citizens' phone numbers, and then automatically sends them WhatsApp messages.

Almost three weeks ago, 147 million voters in the country went to the polls for legislative elections and the first round of the presidential elections.

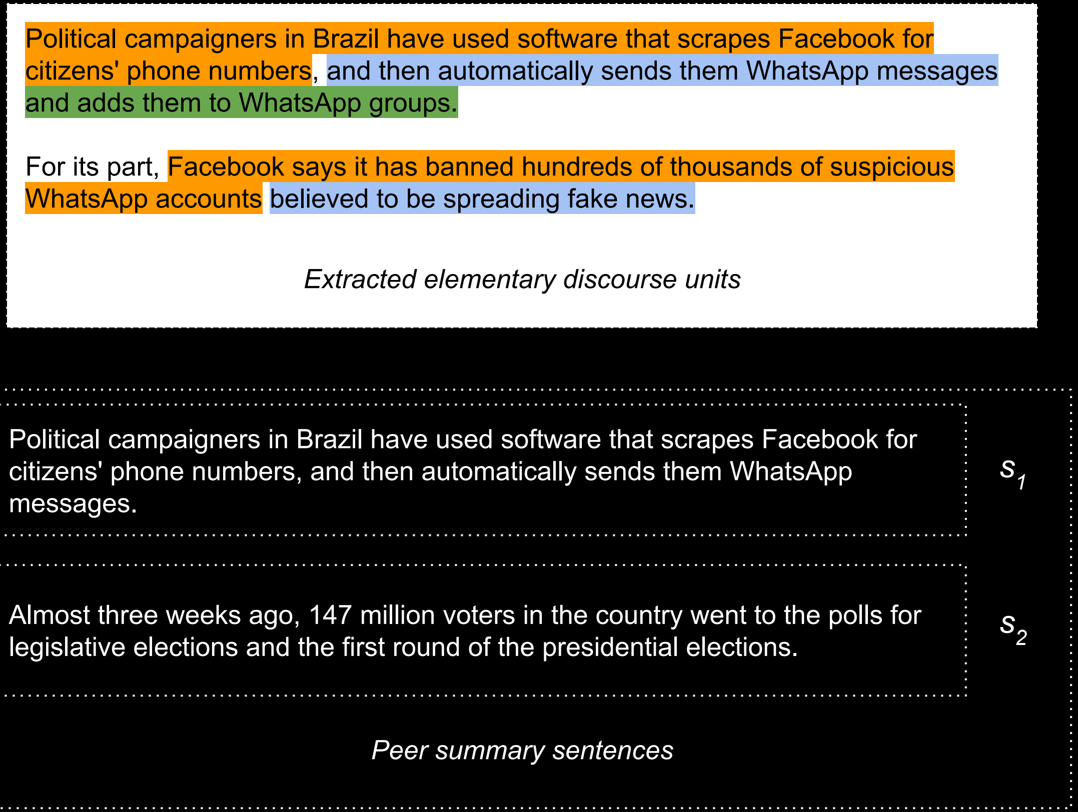
*Peer summary*

# Problem

*(a) Find all peer units that express at least some facts from the model unit and mark them.*

*(b) After all such peer units are marked, think about the whole set of marked peer units and answer the question:*

*(c) “The marked peer units, taken together, express about k% of the meaning expressed by the current model unit”, where k can be equal to 0, 20, 40, 60, 80 and 100.*



Political campaigners in Brazil have used software that scrapes Facebook for citizens' phone numbers, and then automatically sends them WhatsApp messages and adds them to WhatsApp groups.

For its part, Facebook says it has banned hundreds of thousands of suspicious WhatsApp accounts believed to be spreading fake news.

*Extracted elementary discourse units*

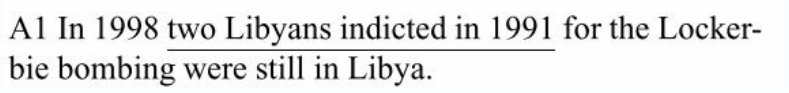
Political campaigners in Brazil have used software that scrapes Facebook for citizens' phone numbers, and then automatically sends them WhatsApp messages. $s_1$

Almost three weeks ago, 147 million voters in the country went to the polls for legislative elections and the first round of the presidential elections. $s_2$

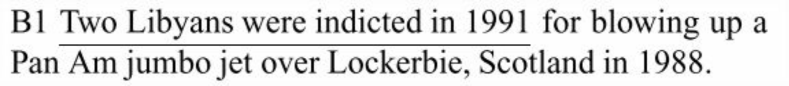
*Peer summary sentences*

# The Pyramid Approach

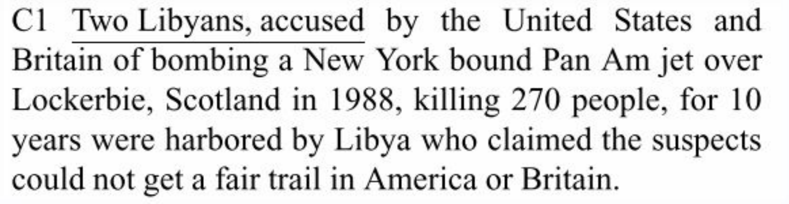
- Approach based on summarization content units (SCUs)
- Starts by identifying similar sentences
- Create SCUs by grouping semantically equivalent sentences
- Weight SCUs by their number of sentences
- Build a pyramid of n tiers where each tier contains SCUs of weight *w*



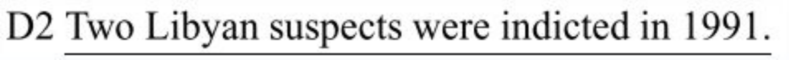
A1 In 1998 two Libyans indicted in 1991 for the Lockerbie bombing were still in Libya.



B1 Two Libyans were indicted in 1991 for blowing up a Pan Am jumbo jet over Lockerbie, Scotland in 1988.



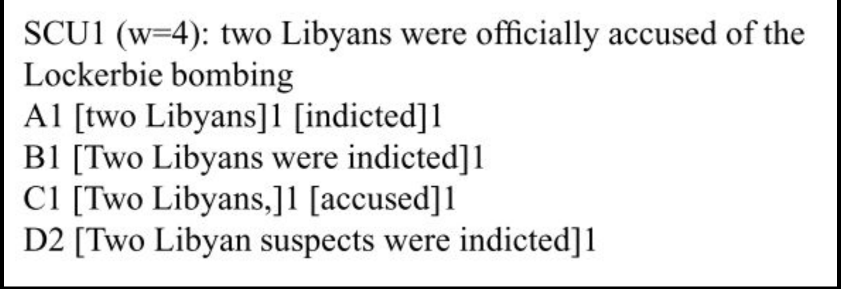
C1 Two Libyans, accused by the United States and Britain of bombing a New York bound Pan Am jet over Lockerbie, Scotland in 1988, killing 270 people, for 10 years were harbored by Libya who claimed the suspects could not get a fair trail in America or Britain.



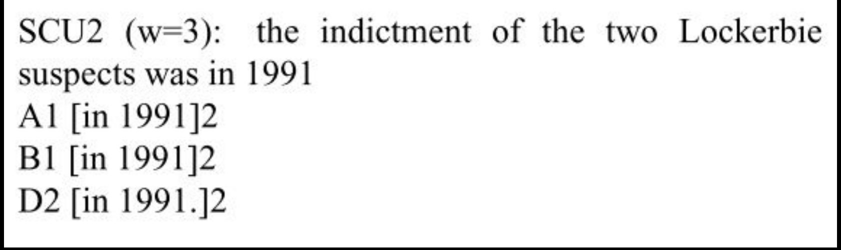
D2 Two Libyan suspects were indicted in 1991.

# The Pyramid Approach

- Approach based on summarization content units (SCUs)
- Starts by identifying similar sentences
- Create SCUs by grouping semantically equivalent sentences
- Weight SCUs by their number of sentences
- Build a pyramid of n tiers where each tier contains SCUs of weight *w*



SCU1 (w=4): two Libyans were officially accused of the Lockerbie bombing
A1 [two Libyans]1 [indicted]1
B1 [Two Libyans were indicted]1
C1 [Two Libyans,]1 [accused]1
D2 [Two Libyan suspects were indicted]1



SCU2 (w=3): the indictment of the two Lockerbie suspects was in 1991
A1 [in 1991]2
B1 [in 1991]2
D2 [in 1991.]2

# The Pyramid Approach

- Approach based on summarization content units (SCUs)
- Starts by identifying similar sentences
- Create SCUs by grouping semantically equivalent sentences
- Weight SCUs by their number of sentences
- Build a pyramid of n tiers where each tier contains SCUs of weight *w*

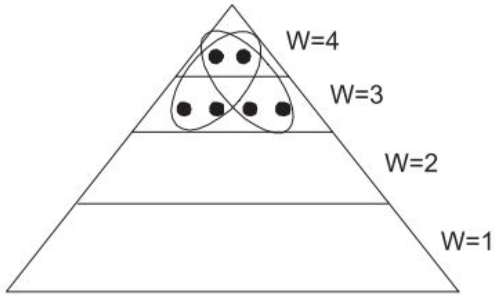
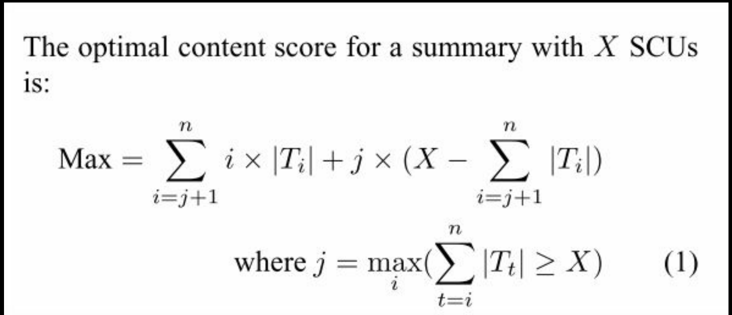


Figure 2: Two of six optimal summaries with 4 SCUs

# The Pyramid score

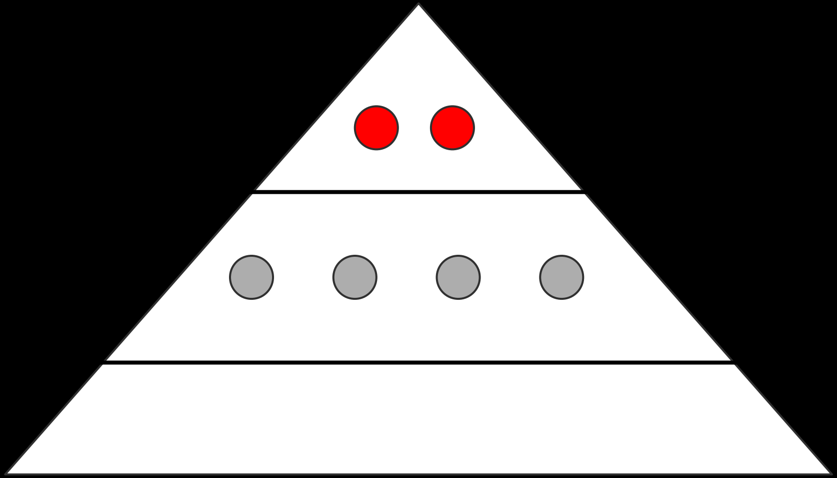
*New summary*



The optimal content score for a summary with $X$ SCUs is:

$$\text{Max} = \sum_{i=j+1}^{n} i \times |T_i| + j \times \left(X - \sum_{i=j+1}^{n} |T_i|\right)$$

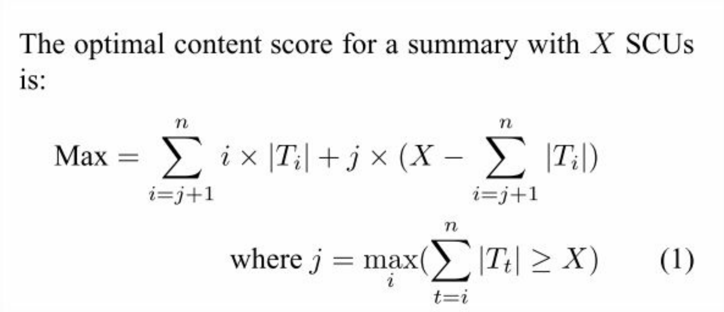
$$\text{where } j = \max_i \left(\sum_{t=i}^{n} |T_t| \geq X\right) \qquad (1)$$



*Optimal score with 2 SCUs*

# The Pyramid score

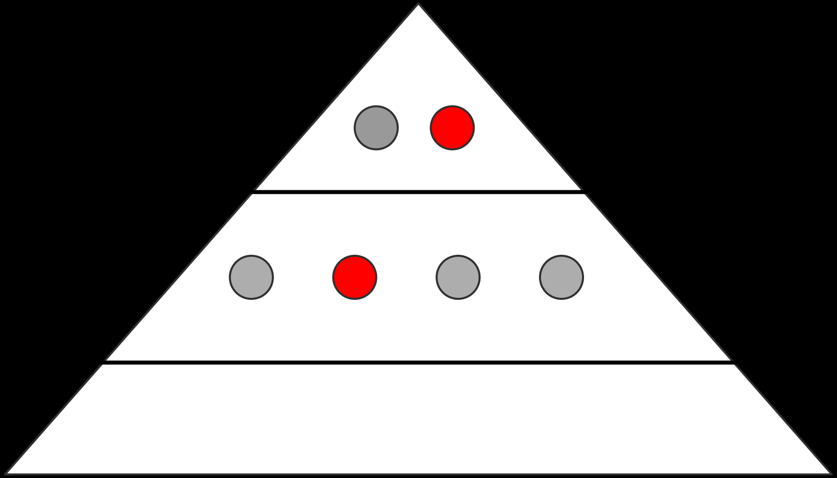
*New summary*



The optimal content score for a summary with $X$ SCUs is:

$$\text{Max} = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|)$$

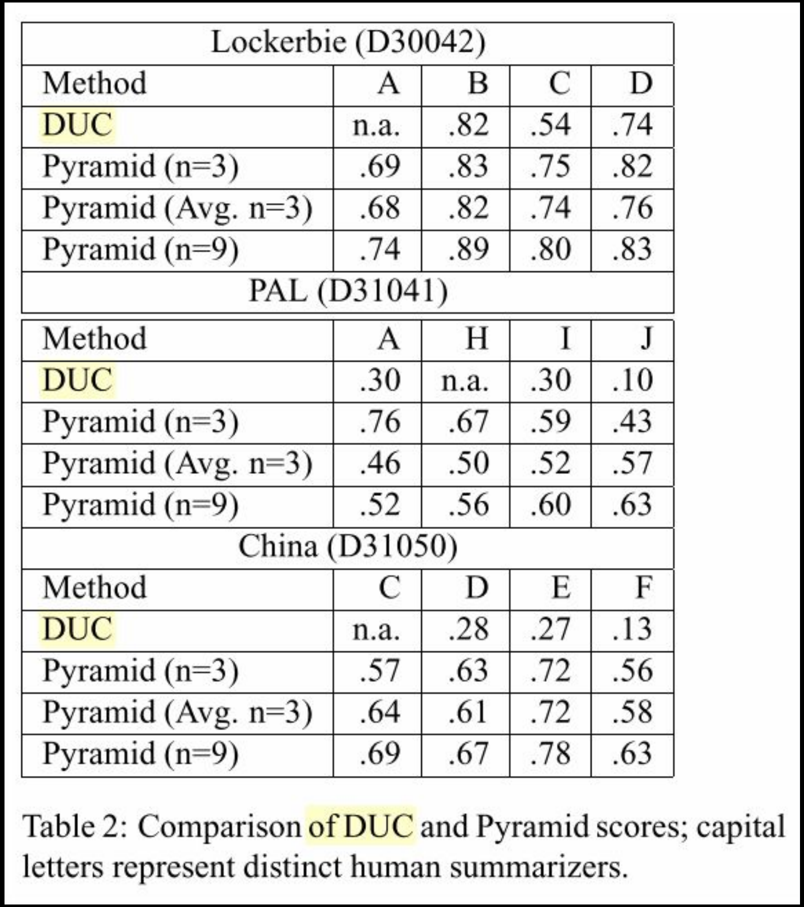
$$\text{where } j = \max_{i}(\sum_{t=i}^{n} |T_t| \geq X) \qquad (1)$$



*Suboptimal score with 2 SCUs*

# Evaluation

- Compare DUC scores of 3 summary sets against pyramid scores



Lockerbie (D30042)

| Method | A | B | C | D |
|---|---|---|---|---|
| DUC | n.a. | .82 | .54 | .74 |
| Pyramid (n=3) | .69 | .83 | .75 | .82 |
| Pyramid (Avg. n=3) | .68 | .82 | .74 | .76 |
| Pyramid (n=9) | .74 | .89 | .80 | .83 |

PAL (D31041)

| Method | A | H | I | J |
|---|---|---|---|---|
| DUC | .30 | n.a. | .30 | .10 |
| Pyramid (n=3) | .76 | .67 | .59 | .43 |
| Pyramid (Avg. n=3) | .46 | .50 | .52 | .57 |
| Pyramid (n=9) | .52 | .56 | .60 | .63 |

China (D31050)

| Method | C | D | E | F |
|---|---|---|---|---|
| DUC | n.a. | .28 | .27 | .13 |
| Pyramid (n=3) | .57 | .63 | .72 | .56 |
| Pyramid (Avg. n=3) | .64 | .61 | .72 | .58 |
| Pyramid (n=9) | .69 | .67 | .78 | .63 |

Table 2: Comparison of DUC and Pyramid scores; capital letters represent distinct human summarizers.

# Reliability and Robustness

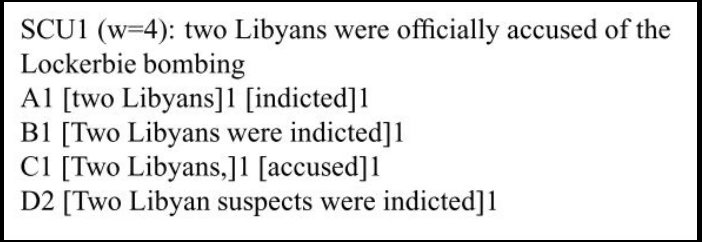
Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

- Treat every word in a summary as a coding unit
- Treat the SCU containing that word as its label
- Assign contributors to SCU
- Compute equivalence classes between contributors coreferencing the same SCU label

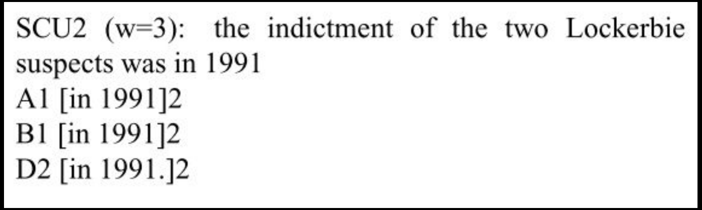
# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

| | A1 | B1 | C1 | D1 |
|---|---|---|---|---|
| In | SCU2 | SCU2 | SCU2 | SCU2 |
| 1991 | SCU2 | SCU2 | SCU2 | SCU2 |
| two | SCU1 | SCU1 | SCU1 | SCU1 |
| Libyans | SCU1 | ... | ... | ... |
| were | SCU1 | .. | ... | |
| officially | SCU1 | .. | | ... |
| accused | SCU1 | .. | | |



SCU1 (w=4): two Libyans were officially accused of the Lockerbie bombing
A1 [two Libyans]1 [indicted]1
B1 [Two Libyans were indicted]1
C1 [Two Libyans,]1 [accused]1
D2 [Two Libyan suspects were indicted]1



SCU2 (w=3): the indictment of the two Lockerbie suspects was in 1991
A1 [in 1991]2
B1 [in 1991]2
D2 [in 1991.]2

# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

| | A1 | B1 | C1 | D1 |
|---|---|---|---|---|
| In | SCU2 | SCU2 | SCU2 | SCU2 |
| 1991 | SCU2 | SCU2 | SCU2 | SCU2 |
| two | SCU1 | SCU1 | SCU1 | SCU1 |
| Libyans | SCU1 | ... | ... | ... |
| were | SCU1 | .. | ... | |
| officially | SCU1 | .. | | ... |
| accused | SCU1 | .. | | |

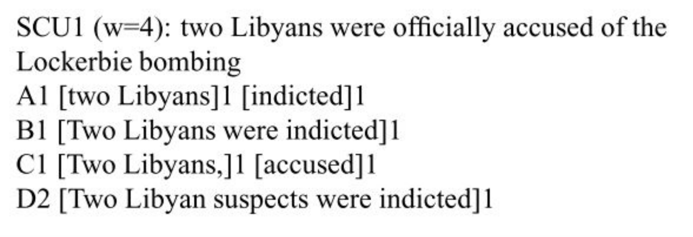
*Looks like perfect agreement*

# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

| | A1 | B1 | C1 | D1 |
|---|---|---|---|---|
| In | SCU2 | SCU2 | SCU2 | SCU2 |
| 1991 | SCU2 | SCU2 | SCU2 | SCU2 |
| two | SCU1 | SCU1 | SCU1 | SCU1 |
| Libyans | SCU1 | ... | ... | ... |
| were | SCU1 | .. | ... | |
| officially | SCU1 | .. | | ... |
| accused | SCU1 | .. | | |

*How to quantify divergences in the words/sentences chosen by A1, B1, C1, D2?*



SCU1 (w=4): two Libyans were officially accused of the Lockerbie bombing
A1 [two Libyans]1 [indicted]1
B1 [Two Libyans were indicted]1
C1 [Two Libyans,]1 [accused]1
D2 [Two Libyan suspects were indicted]1

# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

| | A1 | B1 | C1 | D1 |
|---|---|---|---|---|
| In | ... | ... | ... | ... |
| 1991 | ... | ... | ... | ... |
| two | [two Libyans indicted] | [Two Libyans were indicted] | [Two Libyans accused] | [Two Libyan suspects were indicted] |
| Libyans | [two Libyans indicted] | [Two Libyans were indicted] | [Two Libyans accused] | [Two Libyan suspects were indicted] |
| were | ... | ... | ... | ... |
| officially | ... | ... | ... | ... |

# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

*MASI metric: treats each label of words as a set. Agreement is computed according to set operators*

| | | | | |
|---|---|---|---|---|
| Libyans | [two Libyans indicted] | [Two Libyans were indicted] | [Two Libyans accused] | [Two Libyan suspects were indicted] |
| were | [two Libyans indicted] | [Two Libyans were indicted] | [Two Libyans accused] | [Two Libyan suspects were indicted] |

# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

*MASI metric: treats each label of words as a set. Agreement is computed according to set operators*

| | | | | |
|---|---|---|---|---|
| Libyans | [two Libyans indicted] | [Two Libyans were indicted] | [Two Libyans accused] | [Two Libyan suspects were indicted] |

# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

*MASI metric: treats each label of words as a set. Agreement is computed according to set operators*

| | | | | |
|---|---|---|---|---|
| Libyans | {two Libyans indicted} - {Libyans} | {Two Libyans were indicted} - {Libyans} | ... | ... |

# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

*MASI metric: treats each label of words as a set. Agreement is computed according to set operators*

| | | | | |
|---|---|---|---|---|
| Libyans | {two indicted} | {Two were indicted} | ... | ... |

*Partial agreement*

*MASI = 0.33*

# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

*MASI metric: treats each label of words as a set. Agreement is computed according to set operators*

| | | | | |
|---|---|---|---|---|
| were | [two Libyans indicted] | [Two Libyans were indicted] | [Two Libyans accused] | [Two Libyan suspects were indicted] |

# Reliability and Robustness

Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation

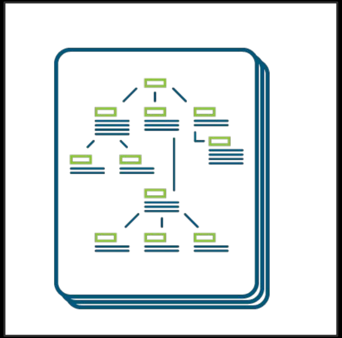
*MASI metric: treats each label of words as a set. Agreement is computed according to set operators*

| | | | | |
|---|---|---|---|---|
| were | {two Libyans indicted} | {Two Libyans indicted} | ... | ... |

*Perfect agreement*

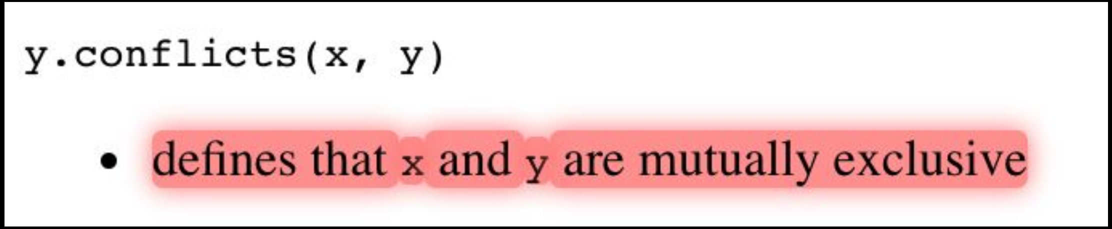
*MASI = 0*

# Why this important to my research?



- I'm studying which are the most relevant pieces of information in software documentation
- I conducted an experiment where participants sought and highlighted relevant information in the documentation of some software resources
- I need to compute their agreement on relevance
- I need to create a pyramid to compare automatic detection techniques against my annotated corpus

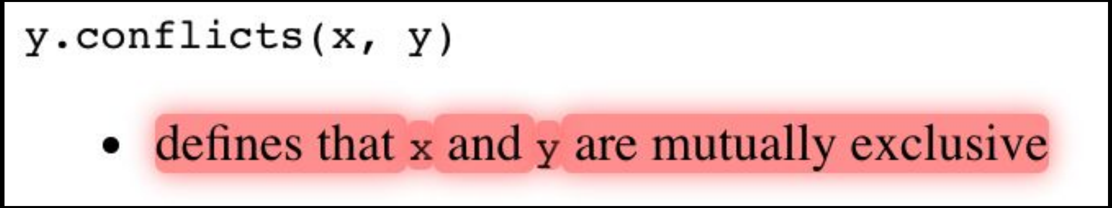
# Why this important to my research?



```
y.conflicts(x, y)
```

- defines that x and y are mutually exclusive

*P20*

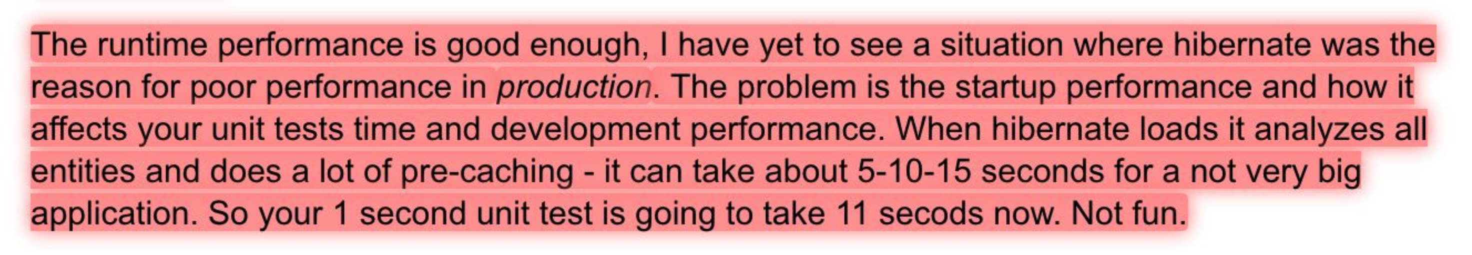


```
y.conflicts(x, y)
```

- defines that x and y are mutually exclusive

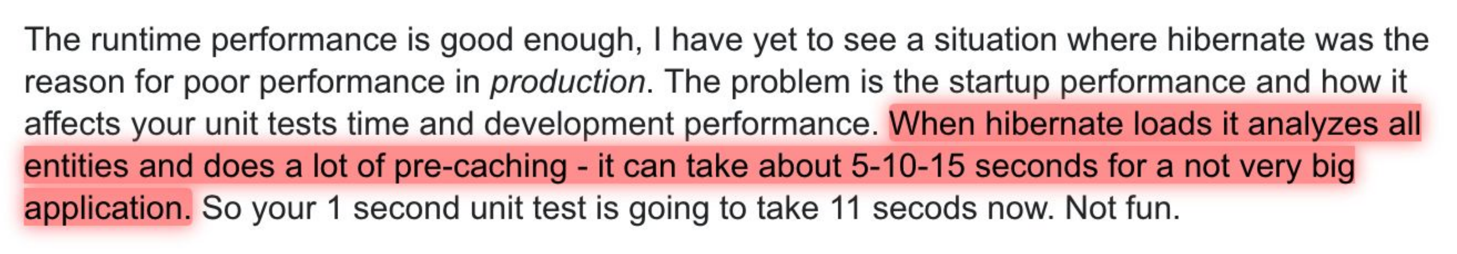
*P03*

# Why this important to my research?



The runtime performance is good enough, I have yet to see a situation where hibernate was the reason for poor performance in *production*. The problem is the startup performance and how it affects your unit tests time and development performance. When hibernate loads it analyzes all entities and does a lot of pre-caching - it can take about 5-10-15 seconds for a not very big application. So your 1 second unit test is going to take 11 secods now. Not fun.

*P20*



The runtime performance is good enough, I have yet to see a situation where hibernate was the reason for poor performance in *production*. The problem is the startup performance and how it affects your unit tests time and development performance. When hibernate loads it analyzes all entities and does a lot of pre-caching - it can take about 5-10-15 seconds for a not very big application. So your 1 second unit test is going to take 11 secods now. Not fun.

*P03*

# Conclusions

Interesting approach to relatively quantify users' subjectivity

Provides robust metrics for computing the quality of automatically produced summaries vs summaries produced by human annotators