Reviewing in NLP – A Tutorial

Kevin Bretonnel Cohen, Karën Fort, Margot Mieskes, Aurélie Névéol

July 5th, 2020











universite



Reviewing in NLP – A Tutorial



Welcome

Kevin Bretonnel Cohen,

UCSOM





Karën Fort,

Sorbonne

Université / Loria

Margot Mieskes, h da Darmstadt

Aurélie Névéol, Université Paris Saclay,

CNRS, LIMSI

Reviewing in NLP – A Tutorial 2020-06-13

Welcome

on behalf of the whole team!

└─Welcome

Why do reviews? Reviewing in NLP – A Tutorial Why do reviews? 2020-06-13 └─Why do reviews? Reviewing in NLP – A Tutorial take a moment to think about this

Why do reviews?

Reviewing in NLP – A Tutorial

Kevin retonnel ien, Karë t, Margo Jieskes,

- For the good of the World
- For the good of the research community
- For your own good

Reviewing in NLP – A Tutorial

└─Why do reviews?

2020-06-13

For the good of the World
 For the good of the research community
 For your own good

Why do reviews?

3

Who is your audience?

Reviewing in NLP – A Tutorial

Kevin Bretonnel Cohen, Karë Fort, Margo Mieskes,

- The authors of the paper:
 - narrative feedback helps authors understand what the paper conveys
 - actionable feedback helps improve the paper
- The Area Chairs / Editor: *itemize major pros and cons* to help reach a decision
 - *performance point of view* helps them evaluate the novelty/correctness of the research
 - deontological point of view helps them evaluate the way the research was conducted

Reviewing in NLP – A Tutorial

Who is your audience?

Who is your audience?

 The Area Chairs / Editor: itemize major pros and cons to help reach a decision

novelty/correctness of the research
decertological point of view helps them evaluate the way
the research was conducted

Perhaps this helps guide answers to the exercise in the next slide?

Karen: I added these, I'm not sure this is what you intended to put here

Exercise

NLP – A Tutorial



You are reviewing for a high-status conference, and for a low-status conference. How should your reviews differ?

Reviewing in NLP – A Tutorial

└─Exercise



Exercise

You are reviewing for a high-status conference, and for a low-status conference. How should your reviews differ?

Trick Question!

Exercise Reviewing in NLP – A Tutorial Exercise 2020-06-13 You are reviewing for a high-status conference, and for a low-status conference How should your reviews differ? └─Exercise NLP – A Tutorial Trick Question! You are reviewing for a high-status conference, and for a low-status conference. How should your reviews differ?

Why?

Roles

Reviewing in NLP – A Tutorial

Kevin Bretonnel Cohen, Karën Fort, Margot Mieskes, Aurélie Névéo

The Gatekeeper



https://www.wired.com/2014/08/lotr-physics/

Reviewing in NLP – A Tutorial

Roles

The Catalogue

Roles

Reviewing in NLP – A Tutorial

Kevin Bretonnel Cohen, Karën Fort, Margot Mieskes, Aurélie Névéo

The Gatekeeper



https://www.wired.com/2014/08/lotr-physics/

Not (only) YOU!

Reviewing in NLP – A Tutorial

Roles

The Cardenger

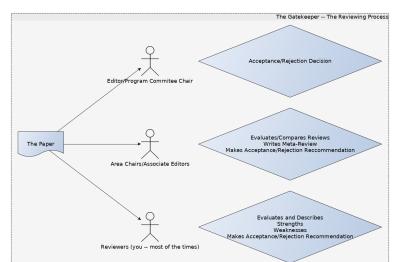
Not (only) YOU!

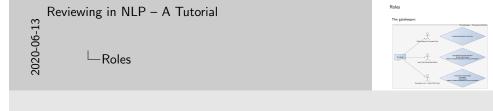
Roles

Reviewing in NLP – A Tutorial

Kevin Bretonnel Johen, Karën Fort, Margot Mieskes,

The gatekeeper:

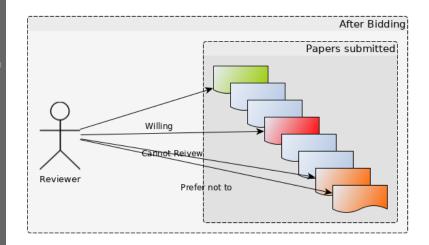




Reviewing in NLP – A Tutorial

Kevin Bretonnel Cohen, Karën Fort, Margot Mieskes,

Bidding



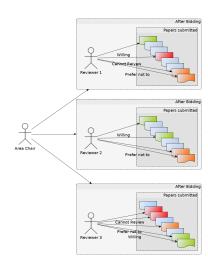
Reviewing in NLP – A Tutorial

 \sqsubseteq How do they get to you?



Reviewing in NLP – A Tutorial

Assignment

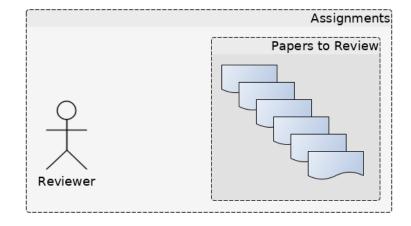


How do they get to you? Reviewing in NLP – A Tutorial 2020-06-13 └─How do they get to you?

Reviewing in NLP – A Tutorial

Kevin Bretonnel Cohen, Karën Fort, Margot Mieskes,

Receiving Assignments



Reviewing in NLP – A Tutorial

└─How do they get to you?

How do they get to you?

Recining Assignments

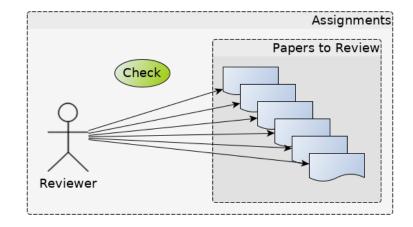
Assignment

Assignm

Reviewing in NLP – A Tutorial

Kevin Bretonnel Cohen, Karën Fort, Margot Mieskes,

Check Papers



Reviewing in NLP – A Tutorial

└─How do they get to you?

How do they get to you?
Clack Pagers
Assignment
Reviewer

Reviewing in NLP – A Tutorial

Kevin Bretonnel Cohen, Karër Fort, Margot Mieskes,

- Conflict of Interest?
- Concerns about your ability?
 - Topic is somehow offensive?
 - Methodology is outside of your skill scope
 - ...

Inform the Area Chair ASAP.



Karen: What about the many many cases where there are no AC (or you don't know them)?

What happens next?

Author Response Period.



https://www.pexels.com/photo/questions-answers-signage-208494/

Reviewing in NLP – A Tutorial

└─What happens next?



What happens next?

What happens next?

Discussion Phase

Kevin Bretonnel Cohen, Karën Fort, Margot Mieskes, Aurélie Névéo

Reviewing in NLP – A Tutorial



Reviewing in NLP – A Tutorial

└─What happens next?



What happens next?

https://freesvg.org/discussion

Reviewing in NLP – A

Kevin Bretonnel Cohen, Karëi Fort, Margoi Mieskes, Aurélie Névés Exercise: Which of the following provides the best support for a conclusion that makes a claim of state-of-the-art performance against a baseline of 0.95?

- Average F-measure as obtained from cross-validation was 0.96.
- Median F-measure as obtained from cross-validation was 0.96.
- Median F-measure as obtained from cross-validation was 0.96 with an inter-quartile range of 0.4, minimum of 0.92, and maximum of 0.96.
- Median F-measure as obtained from cross-validation was 0.94 with an inter-quartile range of 0.02, minimum of 0.91, and maximum of 0.96.

Reviewing in NLP – A Tutorial

Exercise: Which of the following provides the best support for a conclusion that makes a



Average F-measure as obtained from cross-validation was 0.96.

Median F-measure as obtained from

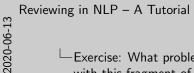
Median F-measure as obtained from cross-validation was 0.95.
Median F-measure as obtained from cross-validation was 0.95 with an inter-quartile range of 0.4, minimum 0.92, and maximum of 0.96.

0.92, and maximum of 0.96.

Median F-measure as obtained fror cross-validation was 0.94 with an inter-quartile range of 0.02, minimu of 0.91, and maximum of 0.96.

Aurélie: Should one of these options state how many experiments the average/median are computed on? Typically, 50 is better than 3...

Point: Summary statistics across multiple trials are valid only with measures of dispersion. Stable performance is preferable to highly variable performance, even with a lower measure of central tendency.





Quantitative score: Methodology: I (Methods are invalid for testing the hypothesis)

Exercise: What problem - if any - do you see with

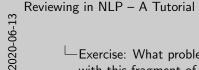
Kevin Bretonnel Cohen, Karër Fort, Margot Mieskes, Aurélie Névéo



Quantitative score: Methodology: 1
(Methods are invalid for testing the hypothesis)

Exercise: What problem – if any – do you see with this fragment of a review?

Exercise: What problem – if any – do you see with this fragment of a review?





Exercise: What problem - if any - do you see with



Comments to authors: The system uses rules. But, machine learning is cheaper than rules.

-Exercise: What problem - if any - do you see with this fragment of a review?

