Кластеризация и сегментация без учителя – зачем и как?

Антон Колонин
akolonin@aigents.com
https://github.com/aigents
Telegram: akolonin

Novosibirsk State University
*THE REAL SCIENCE
https://www.nsu.ru/

AIGENTS
https://aigents.com

AGI IN RUSSIAN
https://agirussia.org

Есть ли у нас план?

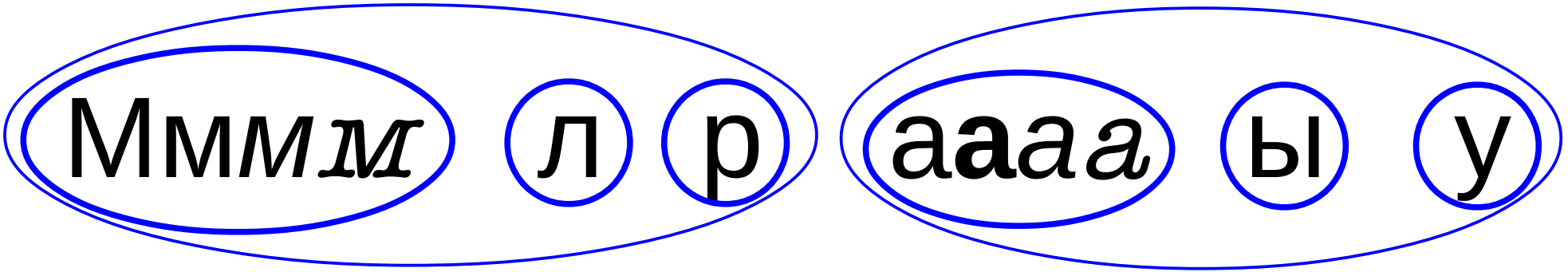
Что такое кластеризация и сегментация?
Зачем они нужны?
- Повышение эффективности обучения с подкреплением
- Прогнозная аналитика на временных рядах
- Интерпретируемая обработка текстов

Кластеризация без учителя – это как?
- Ставим опыты на участниках
- Обсуждаем метрики и результаты эксперимента

Сегментация без учителя – а так можно?
- Загадываем загадку “по теме”
- Обсуждаем метрики/параметры и результаты эксперимента
- Подтверждаем отгадки

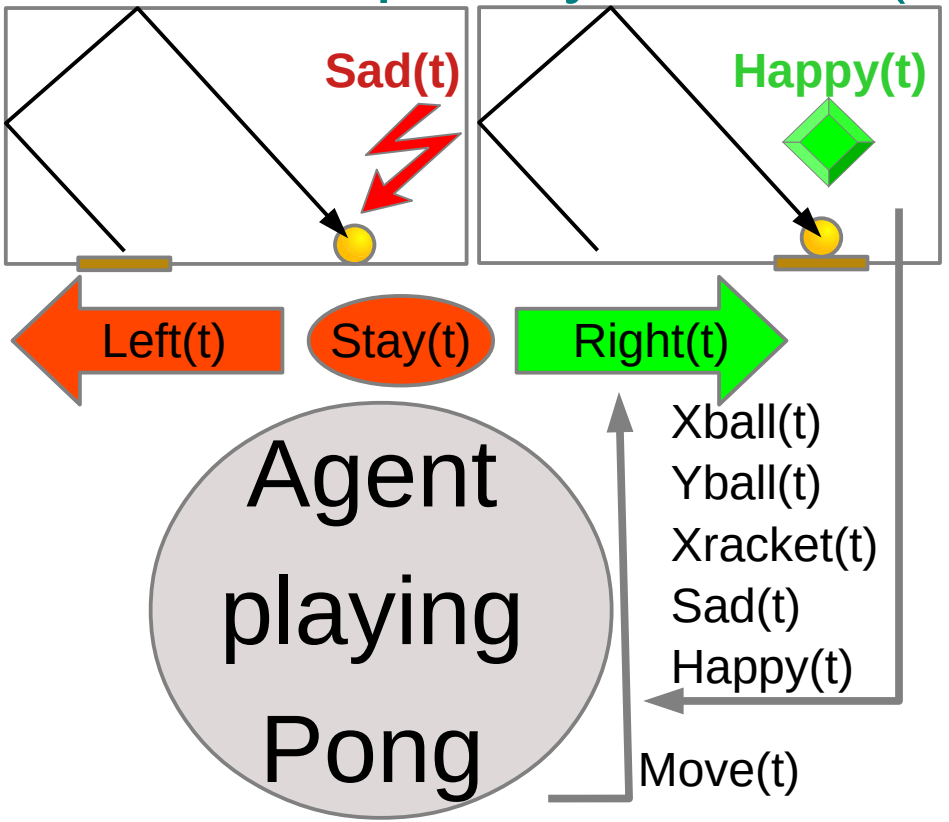
Делаем выводы!

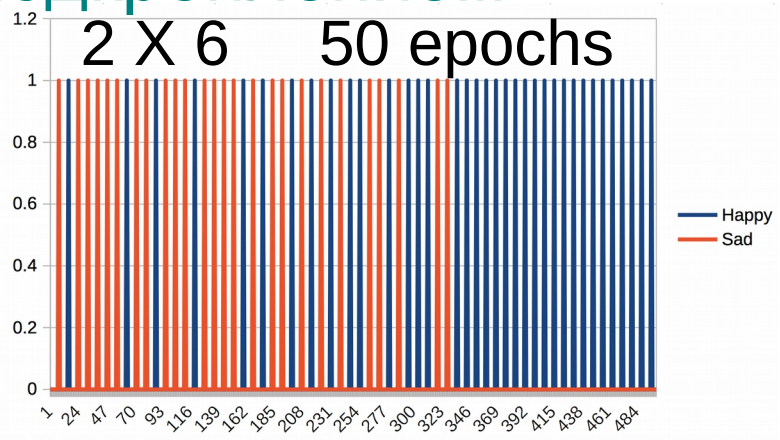
Кластеризация и сегментация

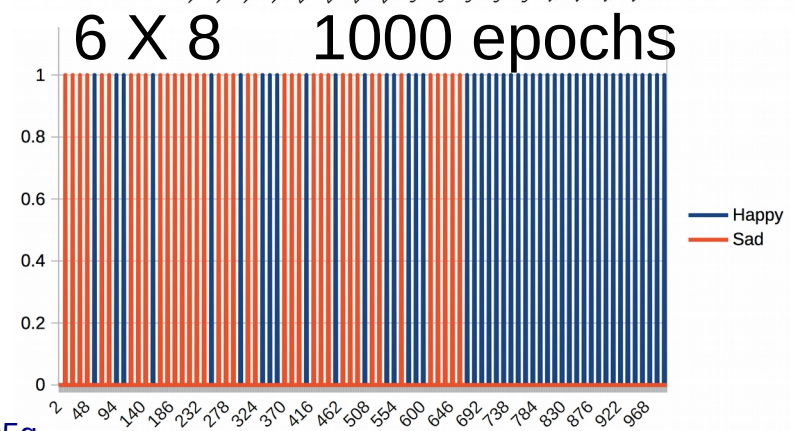


Мама мыла раму

Выявление (не)успешных последовательностей действий при обучении с (само)подкреплением



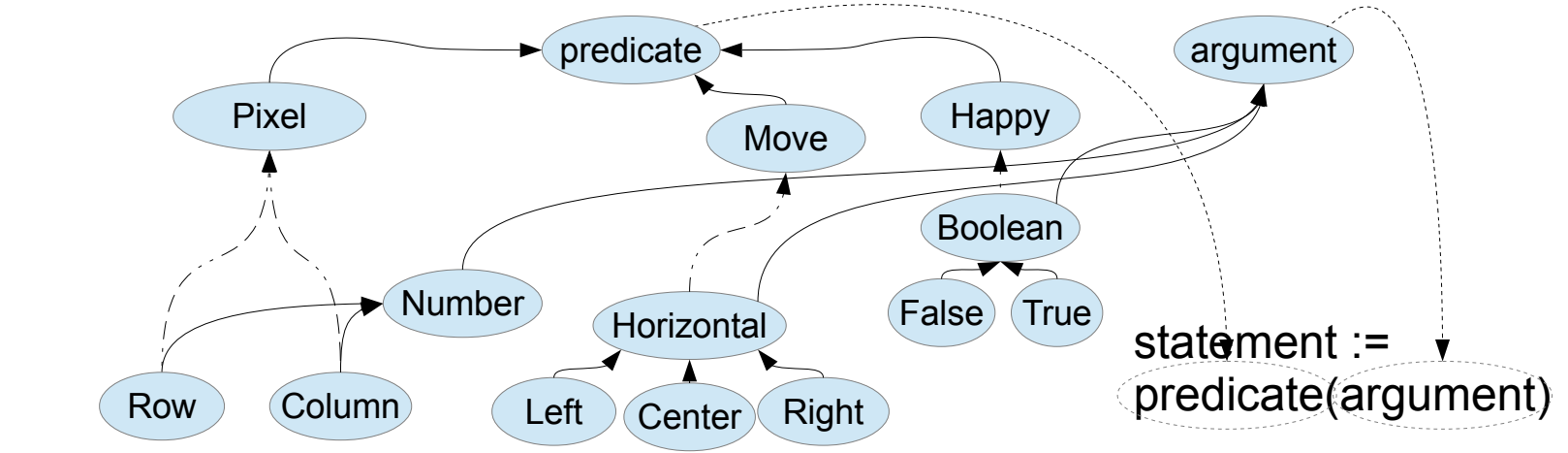


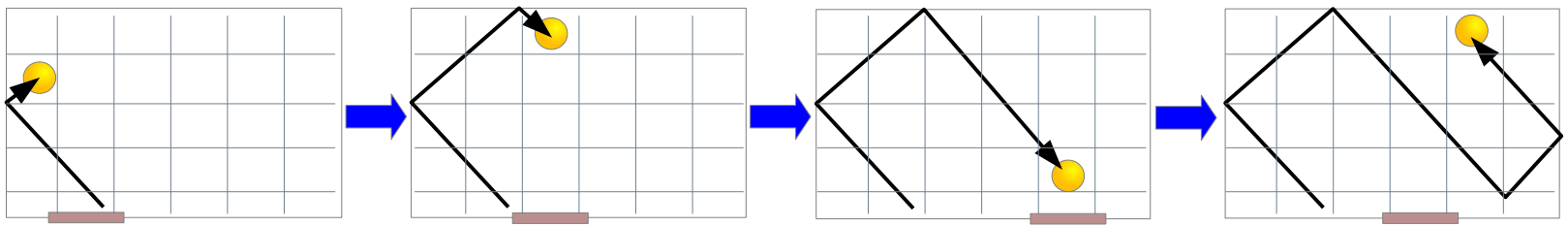


Unsupervised Learning of Temporal Abstractions with Slot-based Transformers
Anand Gopalakrishnan, Kazuki Irie, Jürgen Schmidhuber, Sjoerd van Steenkiste
https://arxiv.org/abs/2203.13573

https://www.youtube.com/watch?v=2LPLhJKh95g
https://www.springerprofessional.de/neuro-symbolic-architecture-for-experiential-learning-in-discret/20008336
https://github.com/aigents/aigents-java/tree/master/src/main/java/net/webstructor/agi

“Дискретное” операционное пространство Pong





Pixel(1,0)
Pixel(4,1)
Happy(False)
=>
Move(Left)

Pixel(0,2)
Pixel(4,2)
Happy(False)
=>
Move(Right)

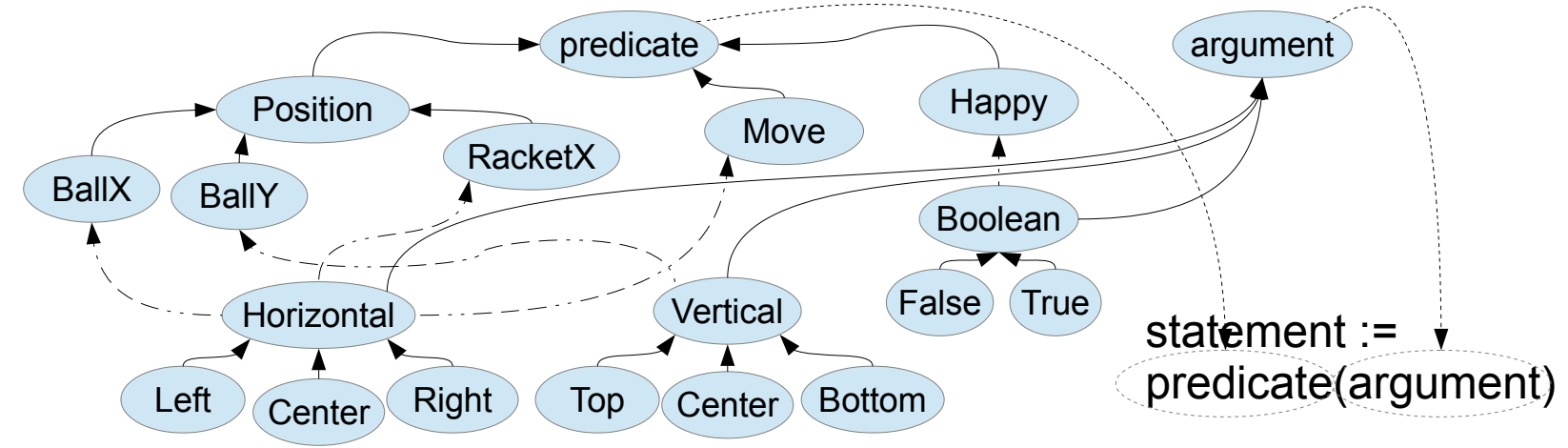
Pixel(3,4)
Pixel(4,4)
Happy(False)
=>
Move(Left)

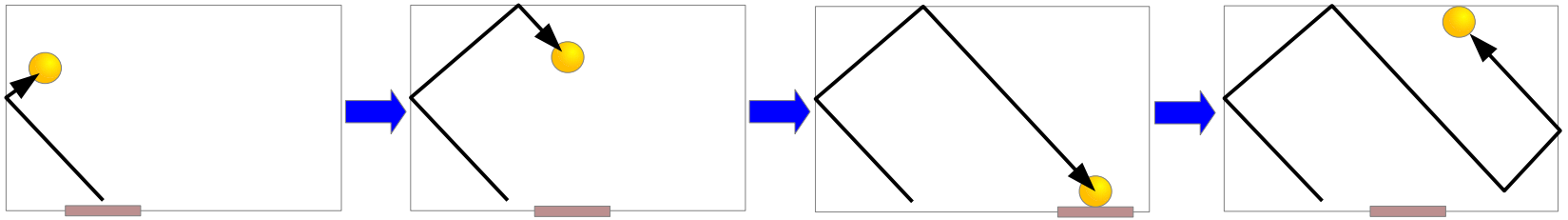
Pixel(0,4)
Pixel(4,3)
Happy(True)
=>
Move(Left)

Global Feedback

https://www.youtube.com/watch?v=2LPLhJKh95g

“Символьное” операционное пространство Pong





BallY(Top)
BallX(Left)
RacketX(Left)
Happy(False)
=> **Move(Left)**

BallY(Top)
BallX(Center)
RacketX(Center)
Happy(False)
=> **Move(Right)**

BallY(Bottom)
BallX(Right)
RacketX(Right)
Happy(False)
=> **Move(Right)**

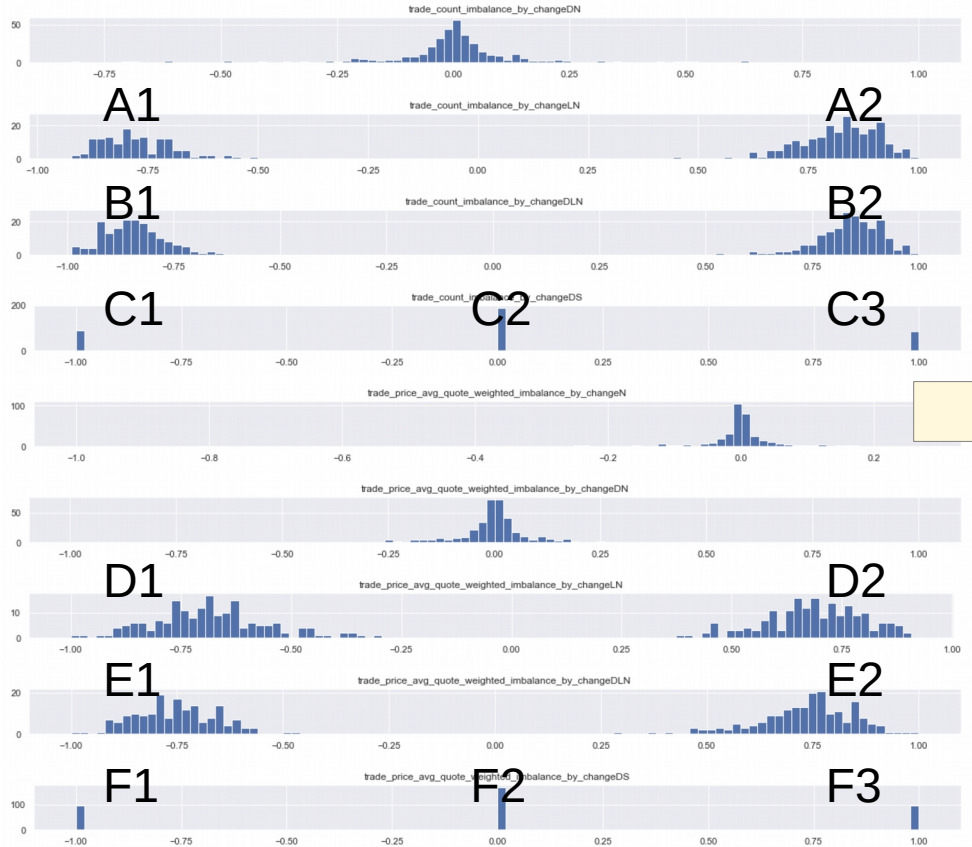
BallY(Bottom)
BallX(Right)
RacketX(Right)
Happy(True)
=> **Move(Left)**

Global Feedback

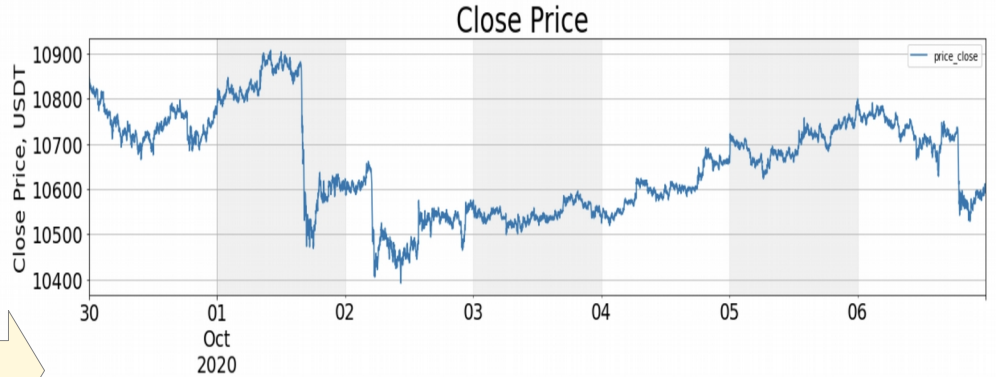
https://www.youtube.com/watch?v=2LPLhJKh95g

“Символизация” состояний рынка

“Квантификация” временных рядов



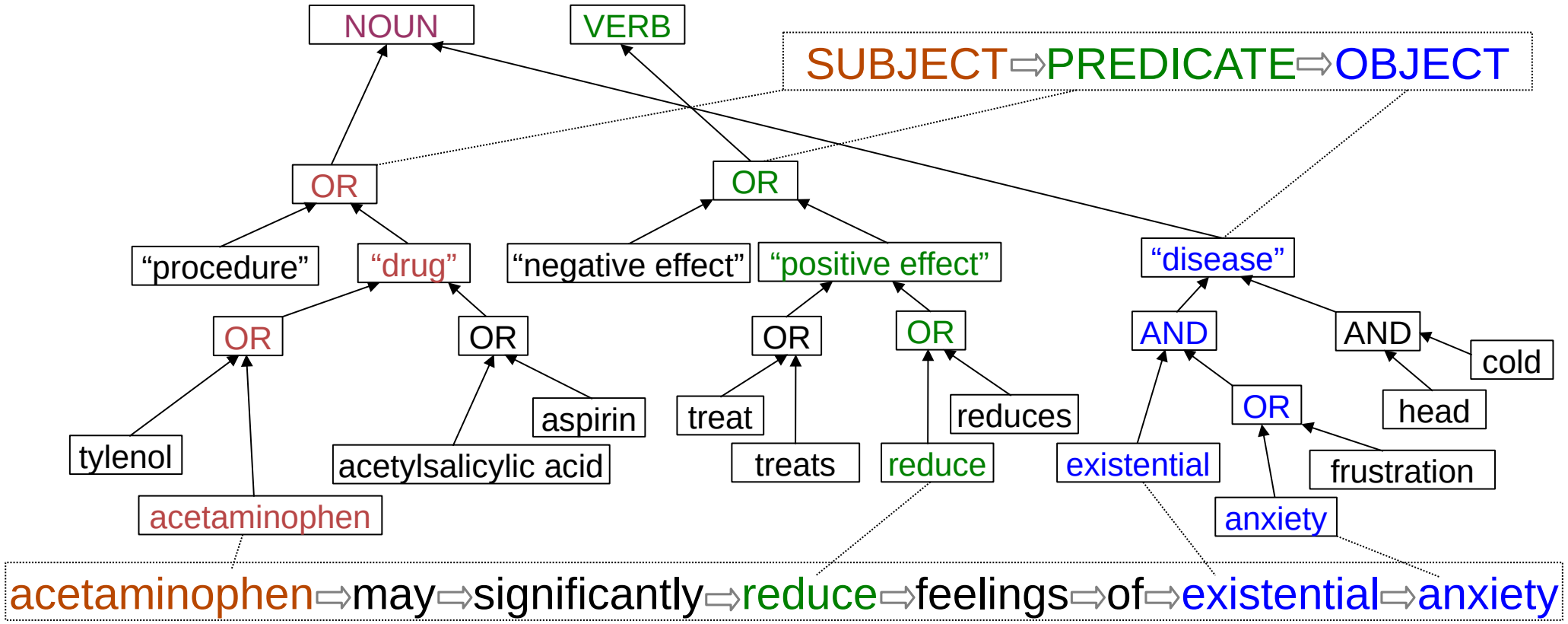
Flat ⇨ Crash ⇨ Crash ⇨ Flat ⇨ Up ⇨ Up ⇨ Crash



A2	A1	A1	A2	A1	A1	A1
B1	B2	B2	B1	B1	B1	B2
C2 ⇨	C1 ⇨	C1 ⇨	C2 ⇨	C3 ⇨	C3 ⇨	C1
D2	D1	D1	D2	D2	D2	D1
E2	E1	E1	E2	E2	E2	E1
F2	F1	F1	F2	F3	F3	F3

Time

Построение гетерархий понятий, отношений, грамматических структур, слов и пунктуации при обучении языку без учителя



https://ieeexplore.ieee.org/document/7361868
https://github.com/aigents/aigents-java
https://github.com/aigents/aigents-java-nlp

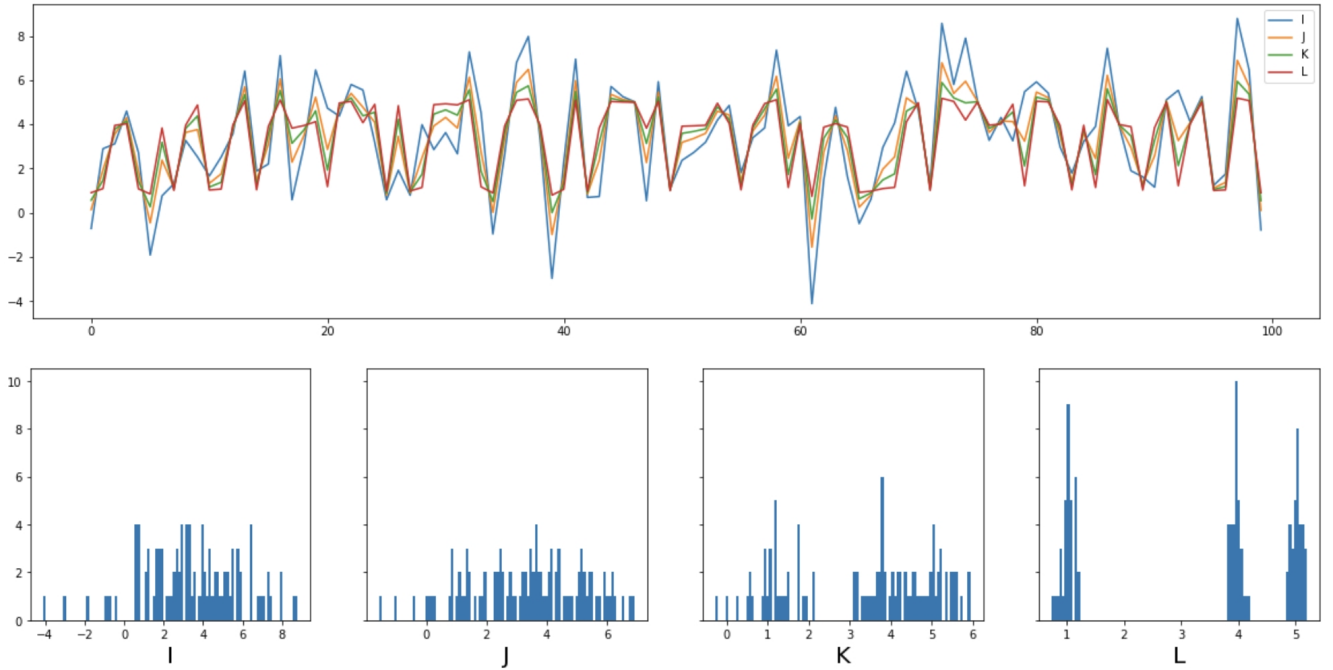
https://www.springerprofessional.de/unsupervised-language-learning-in-opencog/15995030
https://www.springerprofessional.de/en/programmatic-link-grammar-induction-for-unsupervised-language-le/17020348
https://github.com/singnet/language-learning/

Кластеризация без учителя

Не задаем магических значений “K” (требуемого числа кластеров) и порогов для агрегации!

https://github.com/aigents/pygents/blob/main/notebooks/cluster/distribution_modes.ipynb

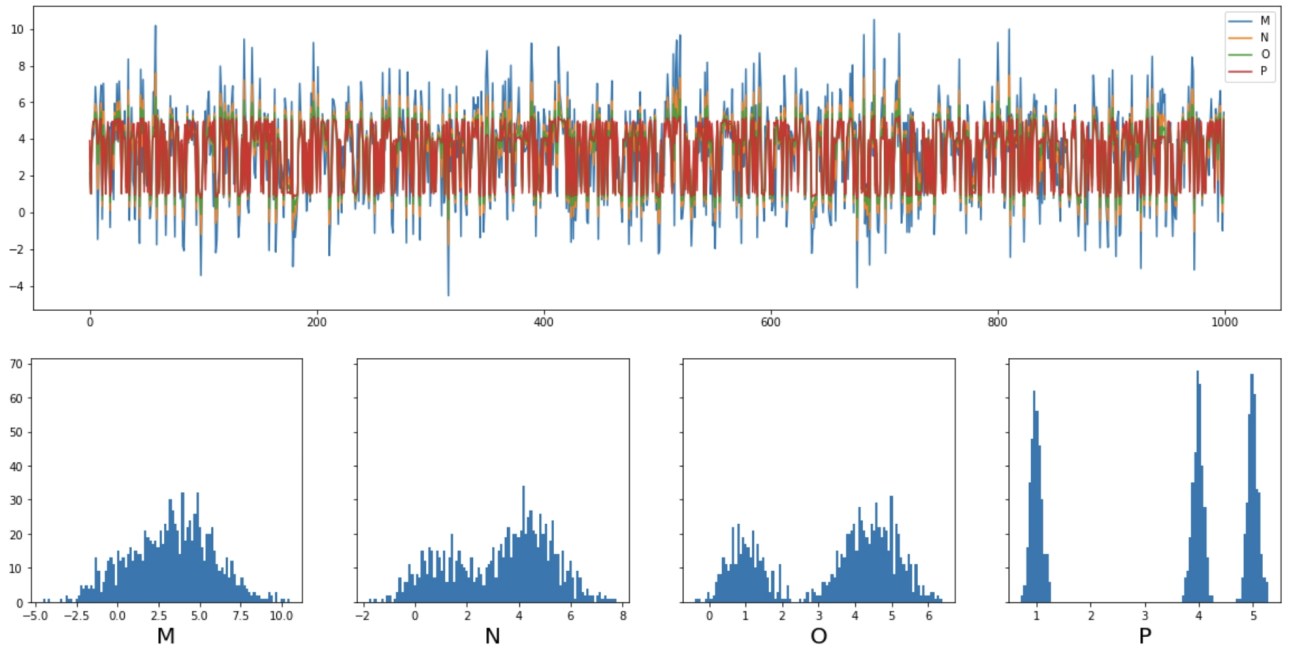
Квантификация – Опыт 1



Human

- 1 or none
- 2
- 3
- 4 or more

Квантификация – Опыт 2



Human

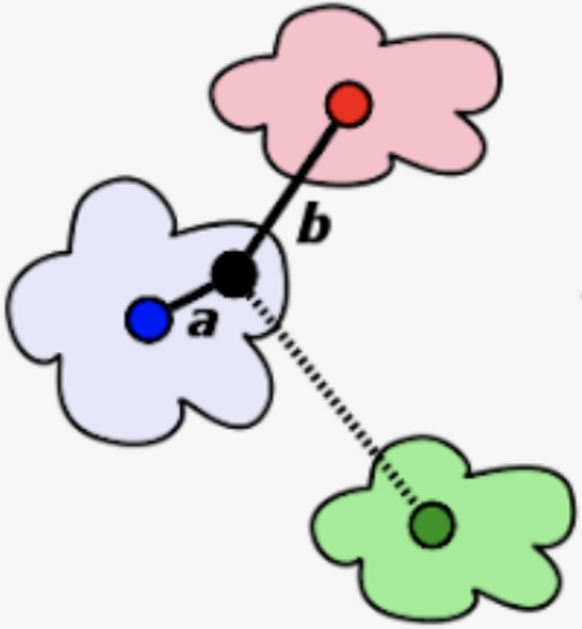
- 1 or none
- 2
- 3
- 4 or more

Метрики кластеризации/квантификации

1. Maximizing "Silhouette Coefficient" (SC), which appears more human-intuitive but does not work for K=1

https://en.wikipedia.org/wiki/Silhouette_(clustering)

https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c



$$SSI_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

http://sujitpal.blogspot.com/2018/03/an-implementation-of-silhouette-score.html

Метрики кластеризации/квантификации

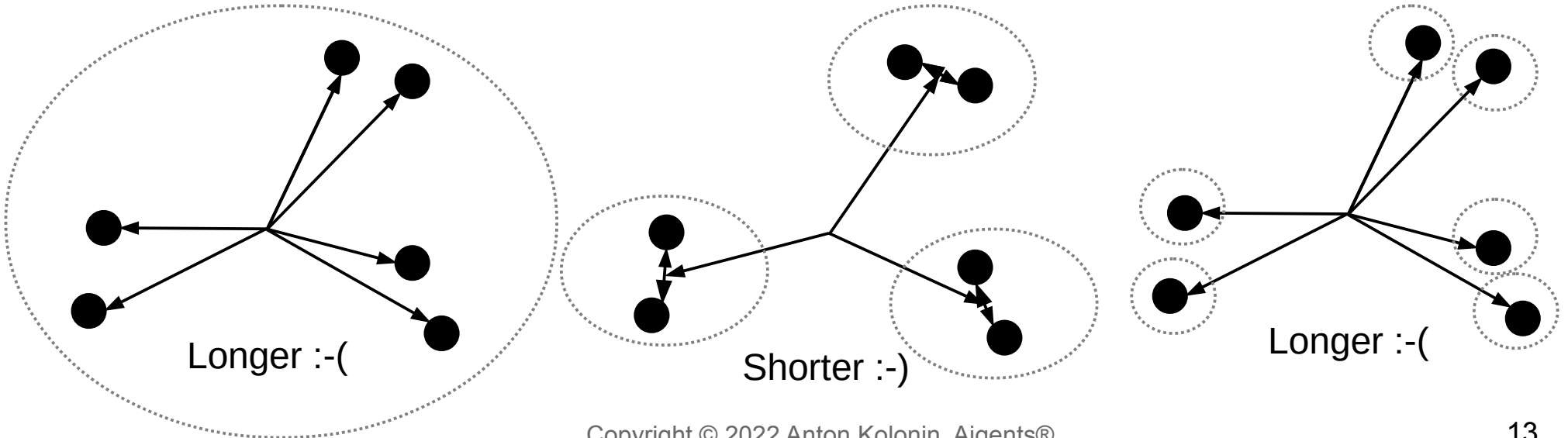
1. Maximizing "Silhouette Coefficient" (SC), which appears more human-intuitive but does not work for K=1

   https://en.wikipedia.org/wiki/Silhouette_(clustering)

   https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

2. Minimizing "Normalized Centroid Distance" (NCD) - based on "minimum description length" idea, works for K=1, does not align with human "reductionist" intuition for diverse distributions (tends to create more clusters than needed)

   https://www.researchgate.net/publication/221020638_Cluster_Validity_Measures_Based_on_the_Minimum_Description_Length_Principle



Метрики кластеризации/квантификации

1. Maximizing "Silhouette Coefficient" (SC), which appears more human-intuitive but does not work for K=1

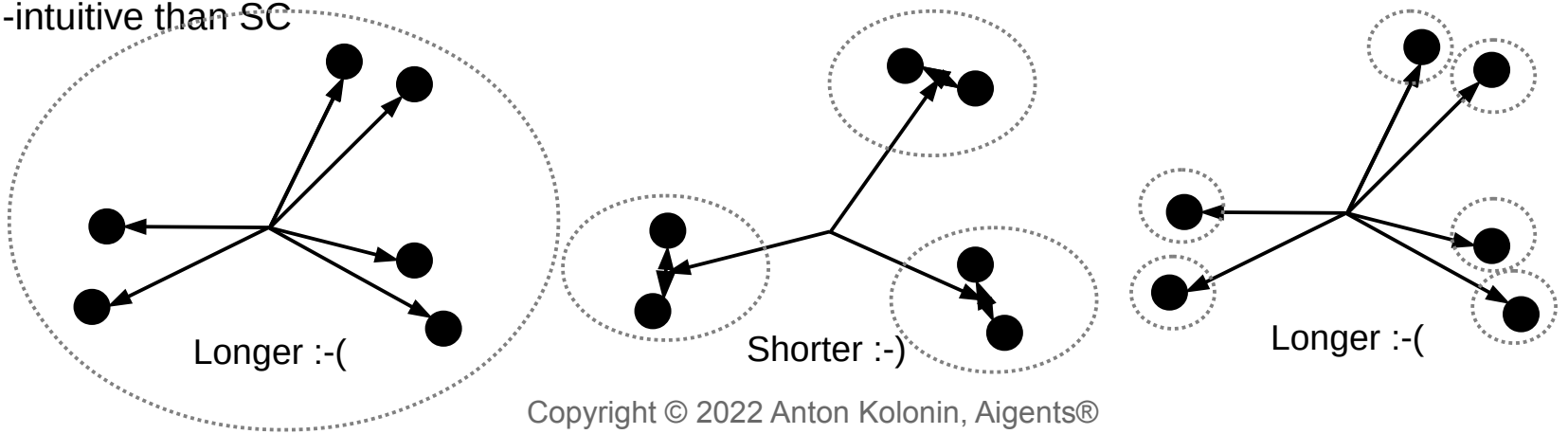
   https://en.wikipedia.org/wiki/Silhouette_(clustering)

   https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

2. Minimizing "Normalized Centroid Distance" (NCD) - based on "minimum description length" idea, works for K=1, does not align with human "reductionist" intuition for diverse distributions (tends to create more clusters than needed)

   https://www.researchgate.net/publication/221020638_Cluster_Validity_Measures_Based_on_the_Minimum_Description_Length_Principle

3. Minimizing "Normalized Centroid Distance times Centroids" (NCDC) - extends NCD multiplying it by number of clusters to penalize creation of too many clusters, works for K=1, more human-intuitive than NCD but less human-intuitive than SC



Метрики кластеризации/квантификации

1. Maximizing "Silhouette Coefficient" (SC), which appears more human-intuitive but does not work for K=1

   https://en.wikipedia.org/wiki/Silhouette_(clustering)

   https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

2. Minimizing "Normalized Centroid Distance" (NCD) - based on "minimum description length" idea, works for K=1, does not align with human "reductionist" intuition for diverse distributions (tends to create more clusters than needed)

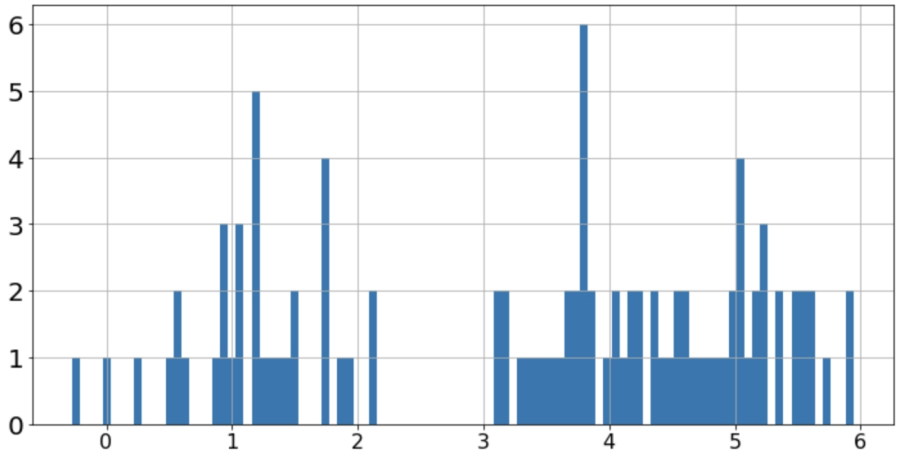
   https://www.researchgate.net/publication/221020638_Cluster_Validity_Measures_Based_on_the_Minimum_Description_Length_Principle

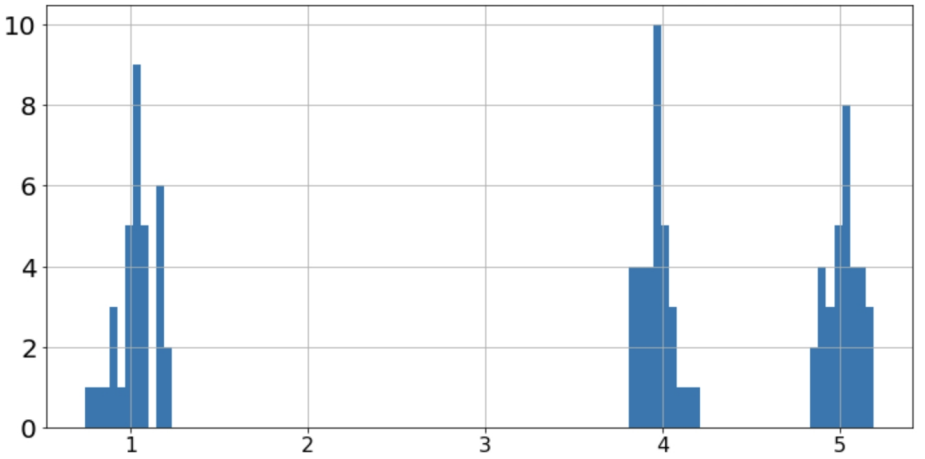
3. Minimizing "Normalized Centroid Distance times Centroids" (NCDC) - extends NCD multiplying it by number of clusters to penalize creation of too many clusters, works for K=1, more human-intuitive than NCD but less human-intuitive than SC

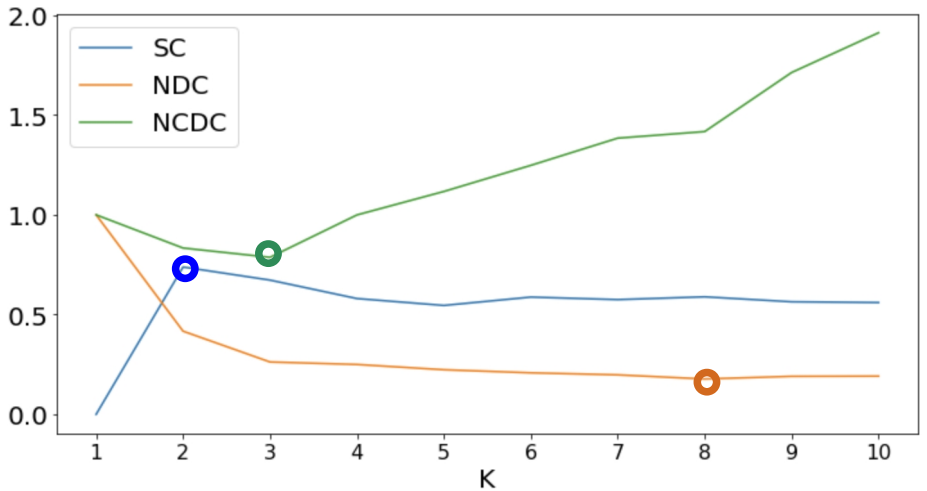
4. Using SC+ (maximize SC if it is above threshold 0.65 or minimize NCDC otherwise) - seems generally optimal from human intuition perspective

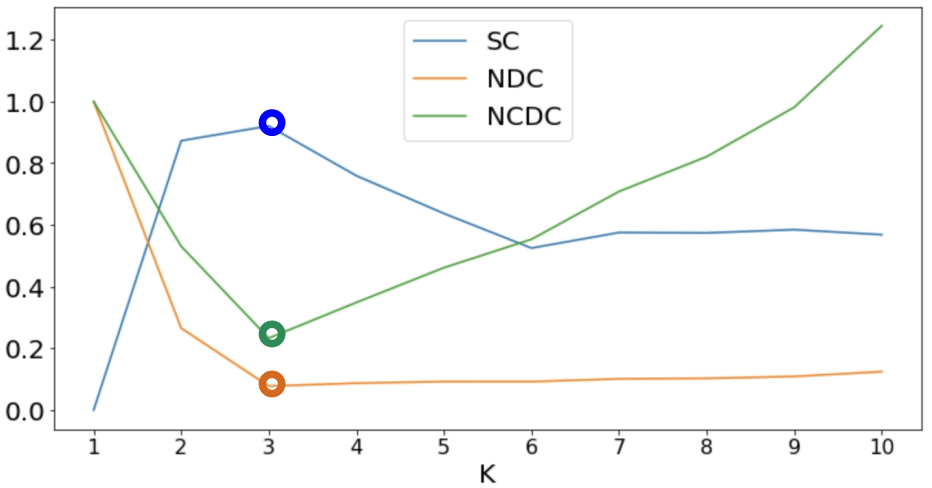
https://github.com/aigents/pygents/blob/main/notebooks/cluster/distribution_modes.ipynb

Метрики квантификации (1-D кластеризации)





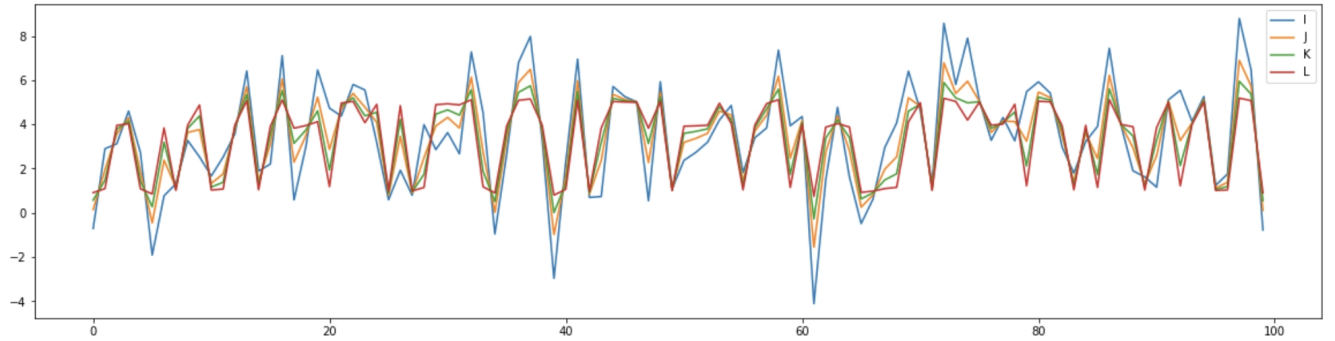




Квантификация – Опыт 1

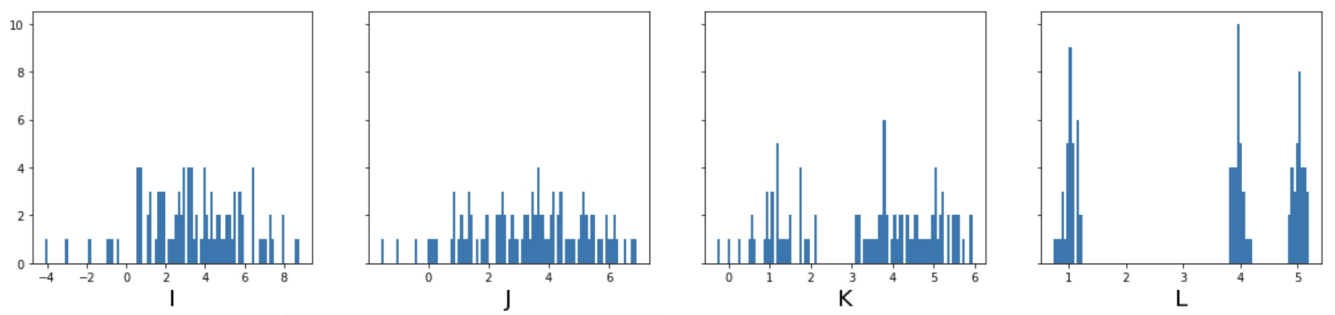
NCDC

I 1
J 1
K 3
L 3



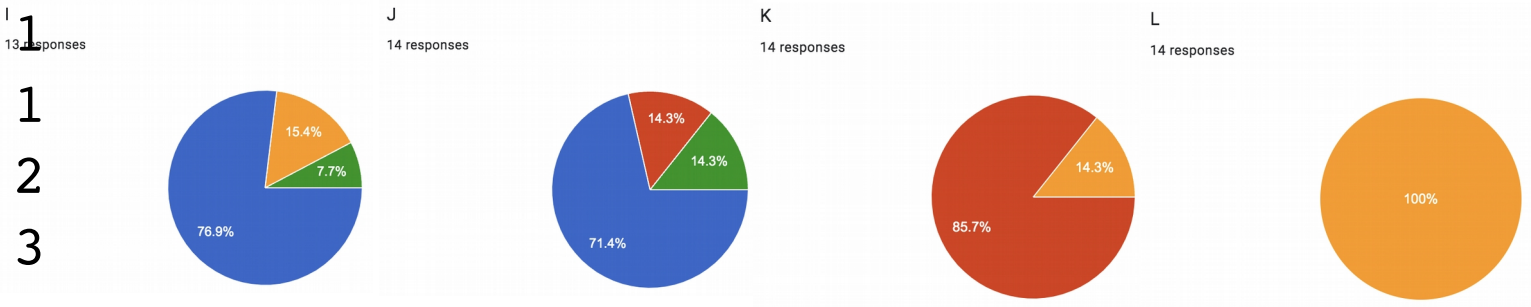
SC+

I 1
J 1
K 2
L 3



Human

I 1
J 1
K 2
L 3



Human

- 1 or none
- 2
- 3
- 4 or more

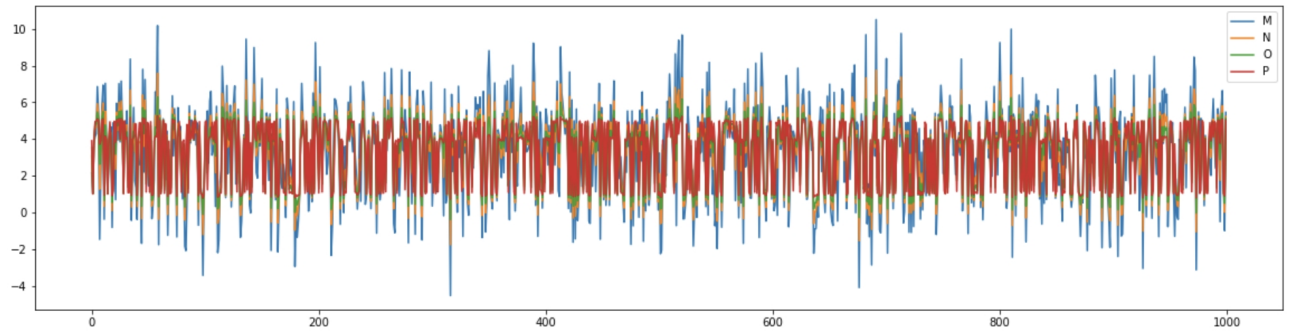
Квантификация – Опыт 2

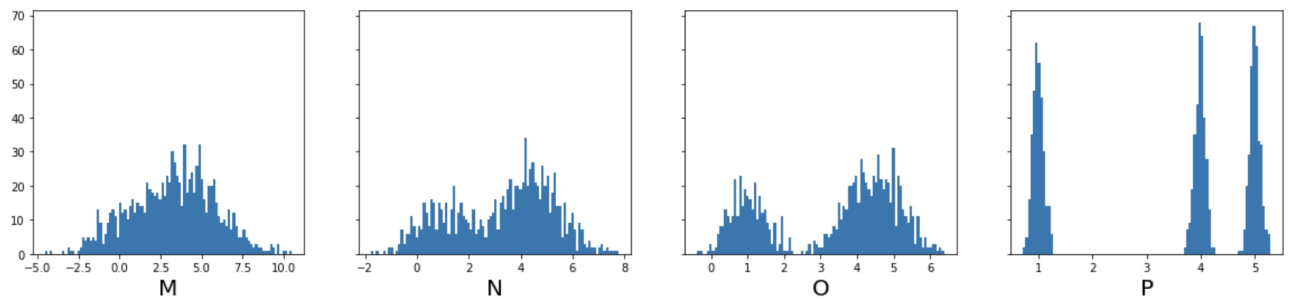
NCDC

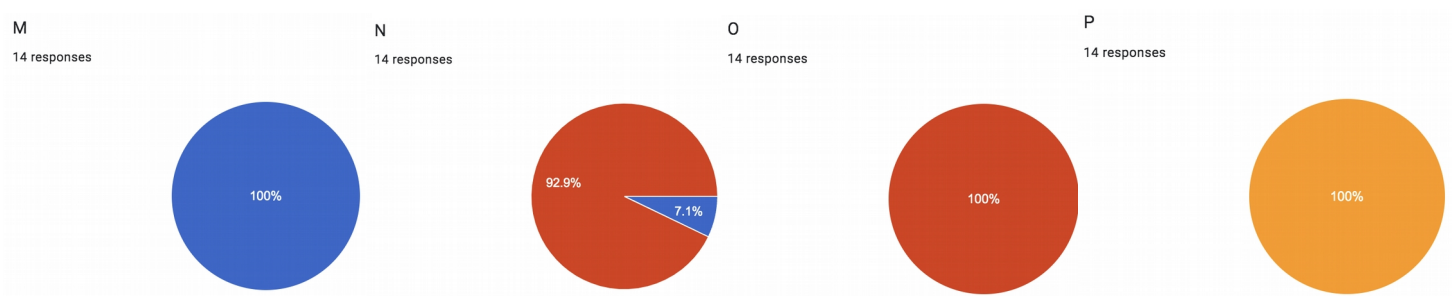
SC+

Human

M 1
N 2
O 2
P 3







Human

- 1 or none
- 2
- 3
- 4 or more

Квантификация: Люди против Машины

Evaluate with "Fleiss' kappa" (FK) and "Krippendorff's alpha" (KA)

https://en.wikipedia.org/wiki/Fleiss%27_kappa

https://stackoverflow.com/questions/51919897/is-fleiss-kappa-a-reliable-measure-for-interannotator-agreement-the-following-r

https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0200-9

https://www.statsmodels.org/dev/generated/statsmodels.stats.inter_rater.fleiss_kappa.html

NCDC vs. SC+:	0.55, 0.56 (Moderate agreement)
Humans vs. humans:	0.59, 0.59 (Moderate agreement)
NCDC vs. humans:	0.47, 0.48 (Moderate agreement)
SC+ vs. humans:	**0.92, 0.92** (Almost perfect agreement)

Сегментация (текста) без учителя

Заранее не знаем ничего о старт/стоп токенах, разделителях, лексиконах и правилах пунктуации!!!

https://github.com/aigents/pygents/tree/main/notebooks/nlp

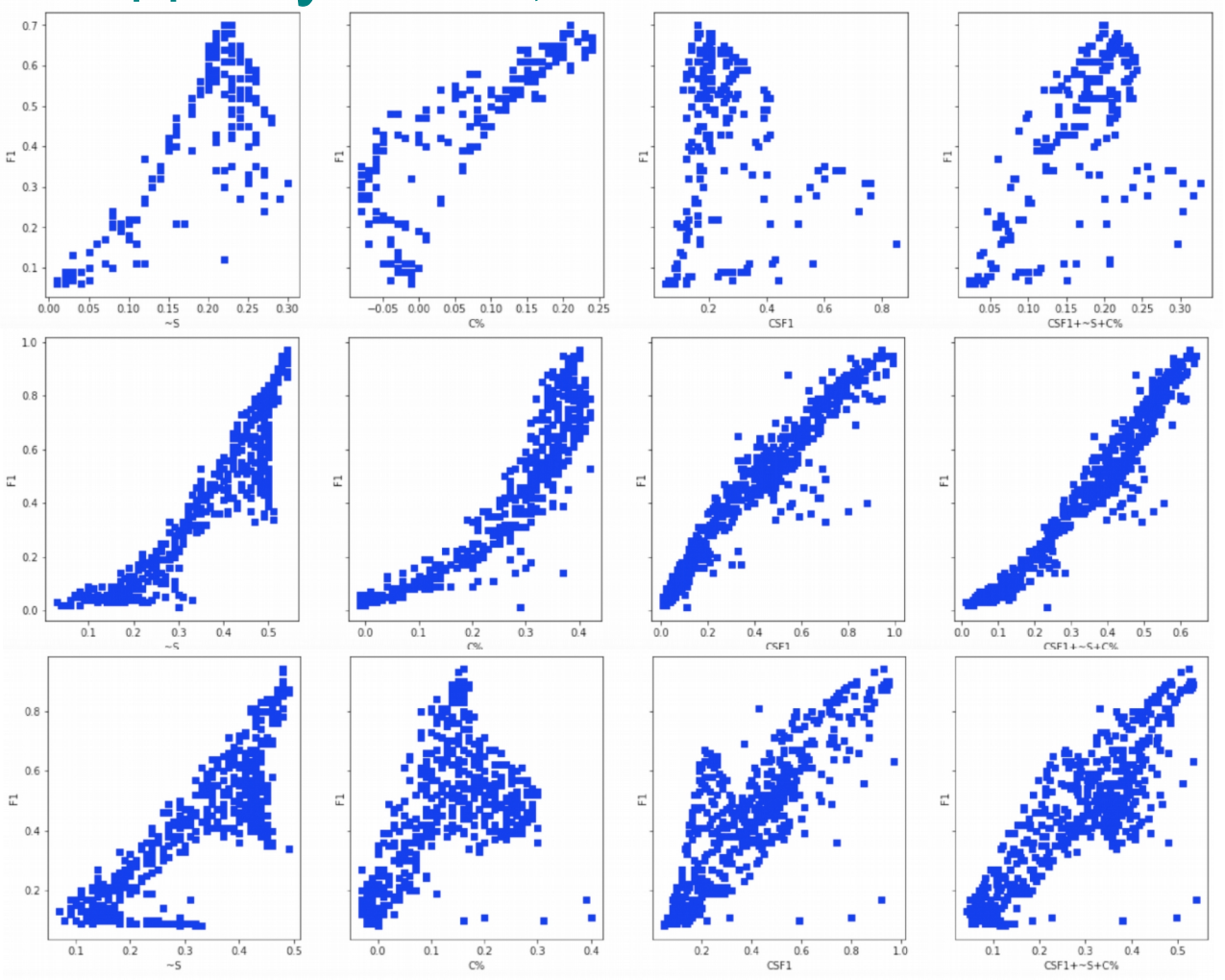
https://arxiv.org/abs/2205.11443

Угадай язык – где Русский, Китайский и Английский?

Язык 1

Язык 2

Язык 3



Токенизация или сегментация текста без учителя – подходы предшественников

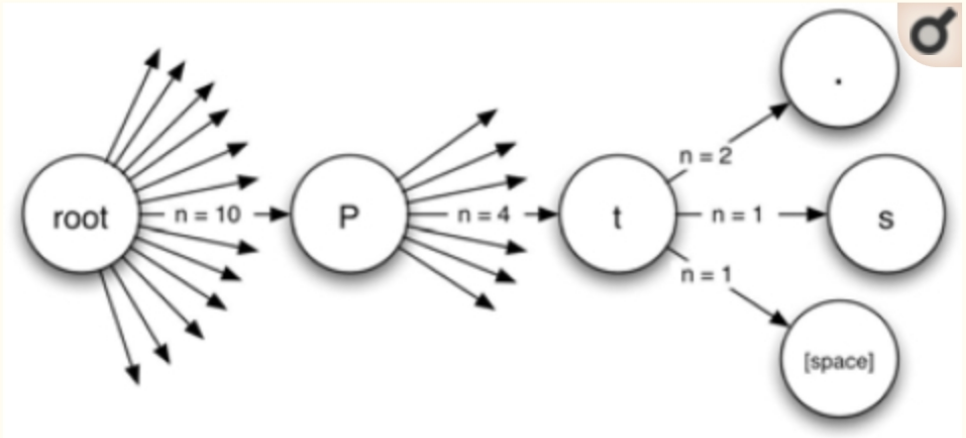


Figure 1

Trie data structure. The probability of observing an ‘s’ given the preceding string “Pt” is ¼, or 25%. The freedom following “pt” is 3.

Metrics/Indicators:

Mutual Information[1]
Conditional Probability[1,2]
Transition Freedom[2,3]

[1] https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=6983&context=etd

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655800/

[3] **Karl Friston.The free-energy principle: a unified brain theory?** https://www.nature.com/articles/nrn2787

Эксперимент – корпуса и методология

Train corpora

Chinese
CLUE News 2016 Validation – 270M
CLUE News 2016 Train – 8,500M

English
Brown – 6M
Gutenberg Children – 29M
Gutenberg Adult – 140M
Social Media – 68M
All above – combined

Russian
RusAge Test – 141M
RusAge Previews – 825M

Test corpus

Parallel Chinese/English/Russian – 100 multi-sentence statements on finance

Metrics/Indicators:

Ngram (Character)
- Probability or Conditional Transition Probability (p-/p+)
  - Deviation (dvp-/dvp+) from mean
  - Derivative (dp-/dp+) and “Peak”
- Transition Freedom (f-/f+)
  - Deviation (dvf-/dvf+) from mean
  - Derivative (df-/df+) and “Peak”

Hyper-parameters:

Combination of Ngram ranks N ([1],[2],[3],[1,2],[1,2,3],…)
Threshold for model compression
Threshold for segmentation

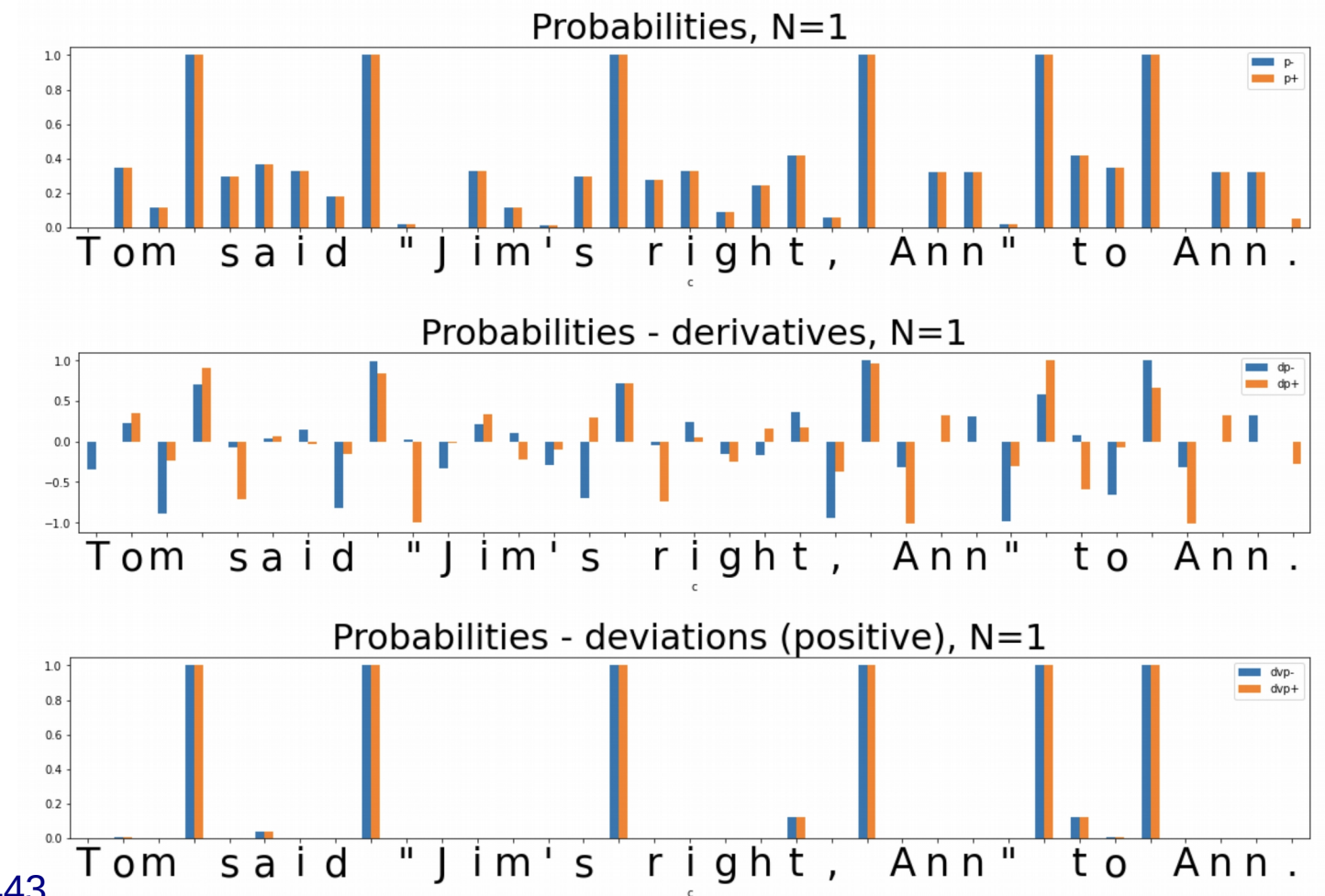
Evaluations:

Tokenization F1, on set of tokens found comparing to delimiter-based (English/Russian) or Jieba (Chineze)
Precision on set of tokens found comparing to reference lexicons

https://arxiv.org/abs/2205.11443

Токеницация по вероятностям токенов/N-gram

Metrics/Indicators:
Ngram (Character)
Probability



https://arxiv.org/abs/2205.11443

Токенизация по условным вероятностям N-gram

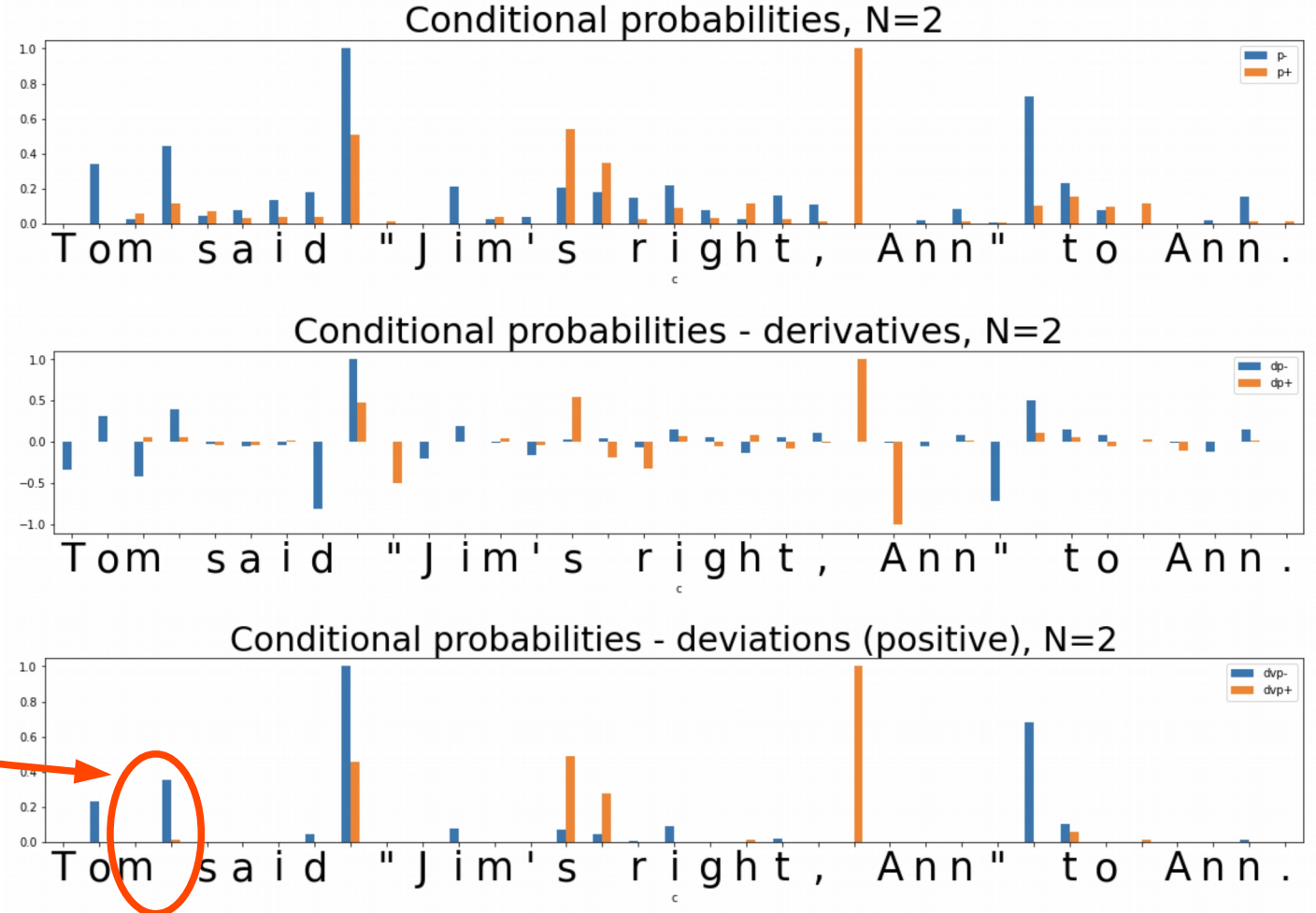
Metrics/Indicators:

Ngram (Character) Conditional Probability

(of Transition)

$P(Ngram_{n+1})/P(Ngram_n)$

P(“m ”)/P(“m”)



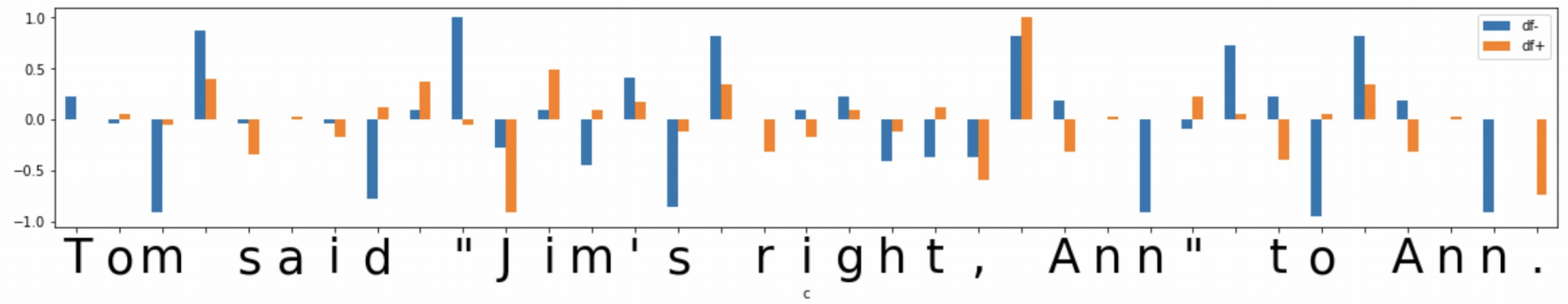
https://arxiv.org/abs/2205.11443

Токенизация по “свободам перехода” (производным)

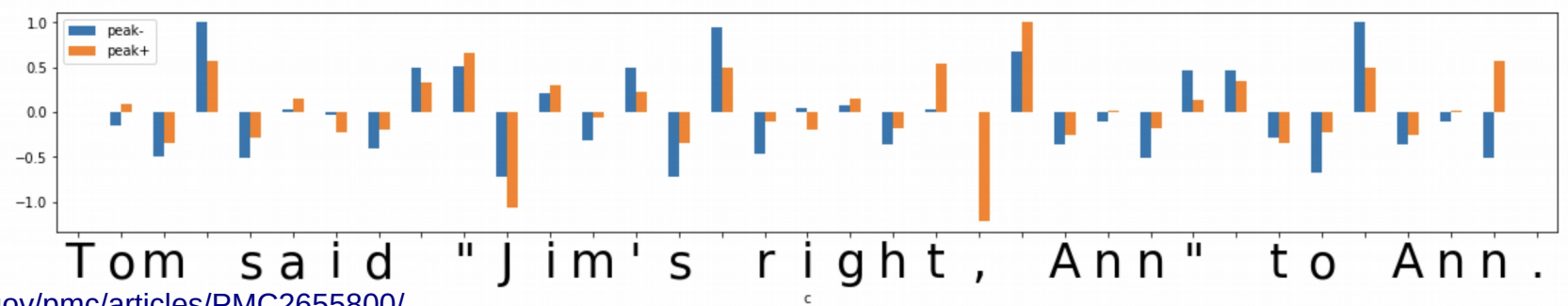
Metrics/ Indicators:

```
Threshold 0.25
Tom said "Jim's right, Ann" to Ann.
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
['Tom', ' ', 'said', ' ', '"', 'Ji', 'm', "'s", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
0.89
Threshold 0.35
Tom said "Jim's right, Ann" to Ann.
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
['Tom', ' ', 'said', ' ', '"', 'Jim', "'s", ' ', 'right', ',', ' ', 'Ann', '"', ' to', ' ', 'Ann', '.']
0.82
```

Transition Freedom (Freedom of Transition) Derivative



Transition Freedom “Peak”)



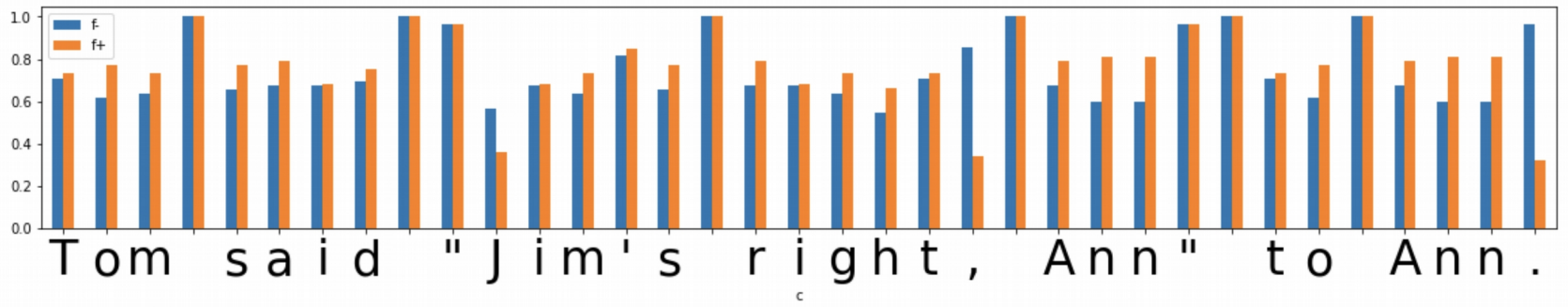
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655800/

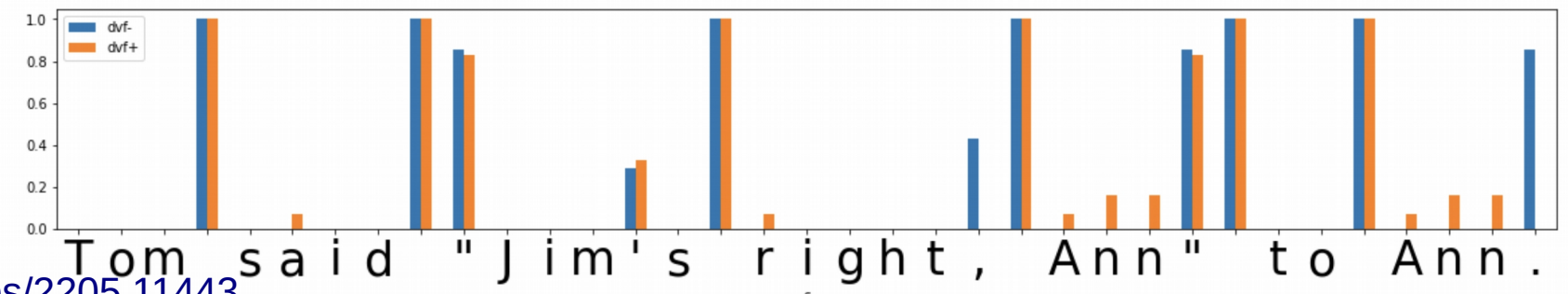
Токенизация по “свободам перехода” (отклонениям)

Metrics/ Indicators:
Transition Freedom (Freedom of Transition)

Transition Freedom Deviation

```
Threshold 0.25
Tom said "Jim's right, Ann" to Ann.
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
['Tom', ' ', 'said', ' ', '"', 'Jim', "'", 's', ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
0.89
Threshold 0.35
Tom said "Jim's right, Ann" to Ann.
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
['Tom', ' ', 'said', ' ', '"', "Jim's", ' ', 'right', ',', ' ', 'Ann', '"', ' ', 'to', ' ', 'Ann', '.']
1.0
```





https://arxiv.org/abs/2205.11443

Токенизация – поиск гиперпарамтров

English

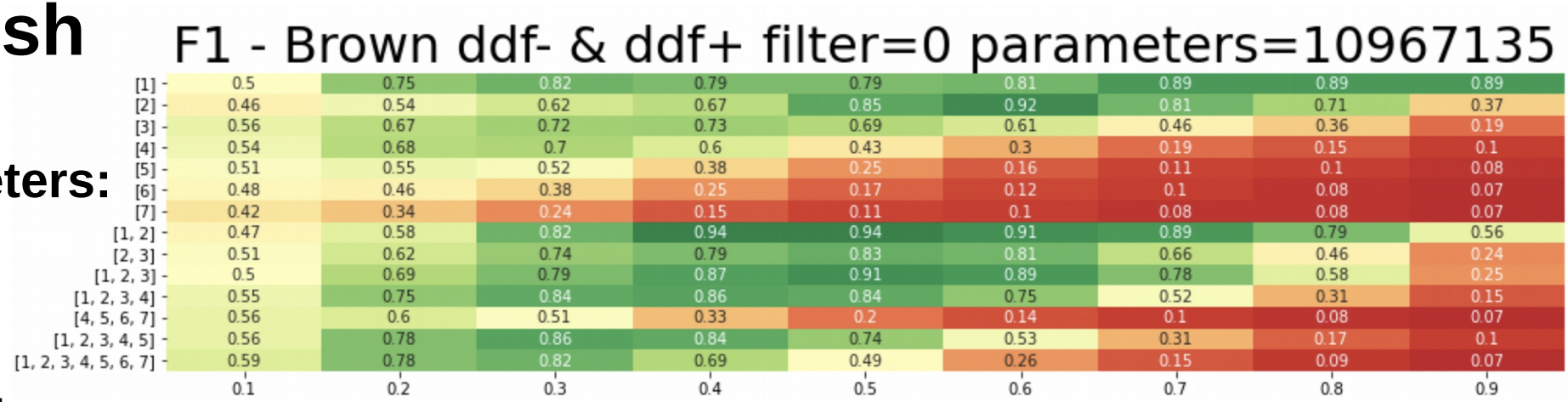
Hyper-Parameters:

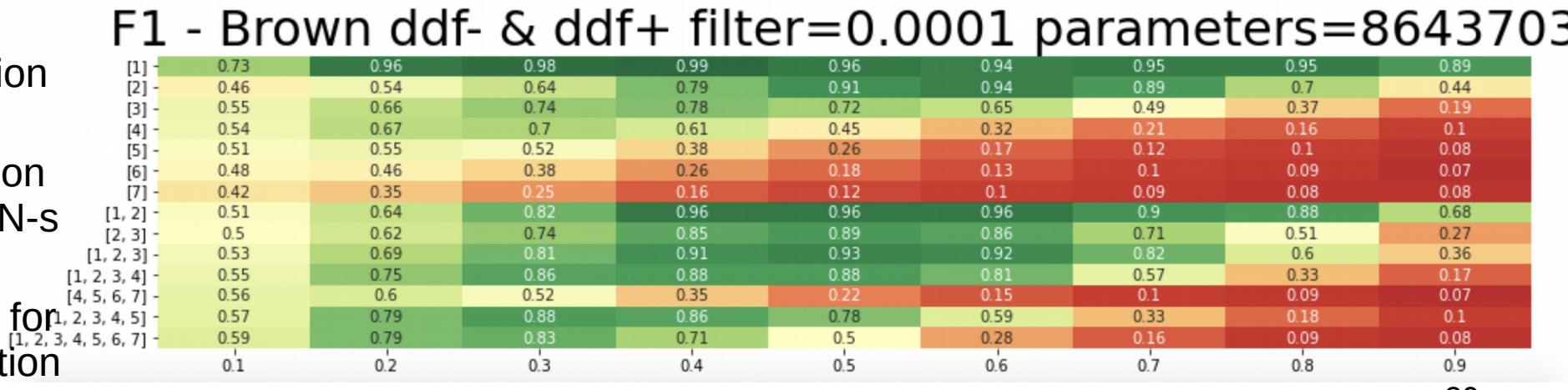
Metric: Transition Freedom

Threshold for model compression

Combination of Ngram N-s

Threshold for segmentation



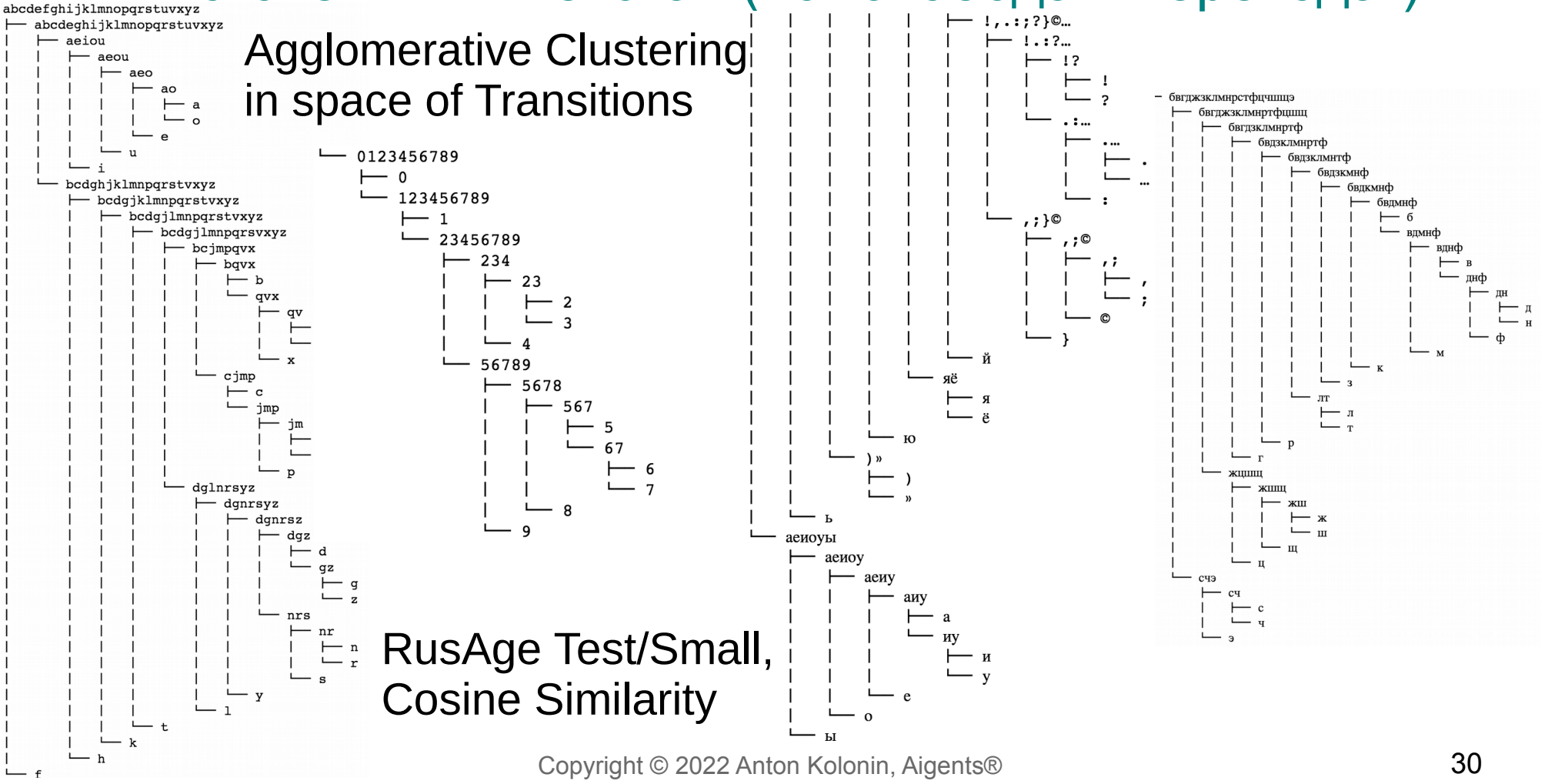


Оценки точности токенизации и обнаружения лексикона

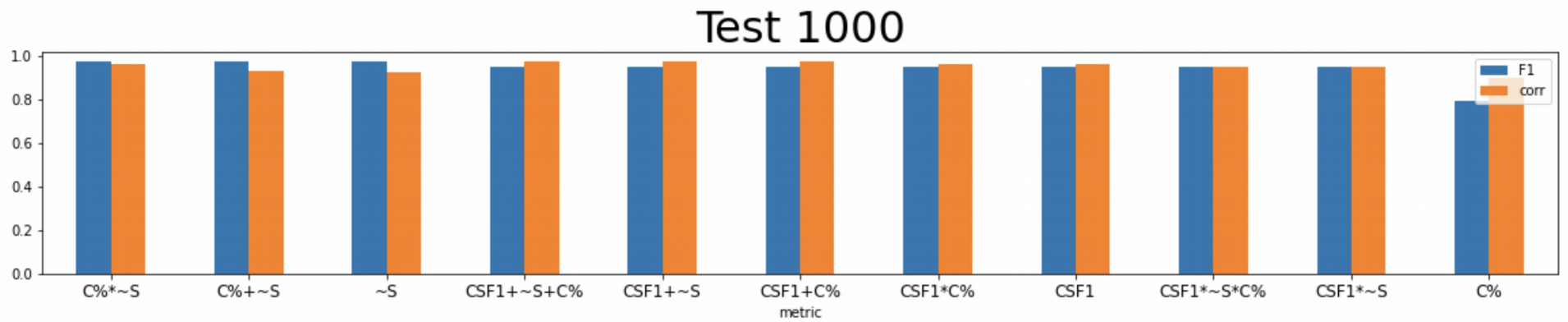
Language	Tokenizer	Tokenization F1	Lexicon Discovery Precision
English	Freedom-based	**0.99**	**0.99** (vs 1.0)
English	Lexicon-based	0.99	-
English no spaces	Freedom-based	0.42	-
English no spaces	Lexicon-based	0.79	-
Russian	Freedom-based	**1.0**	**1.0** (vs 1.0)
Russian	Lexicon-based	0.94	-
Russian no spaces	Freedom-based	0.26	-
Russian no spaces	Lexicon-based	0.72	-
Chinese	Freedom-based	**0.71**	**0.92** (vs 0.94)
Chinese	Lexicon-based	0.83	-

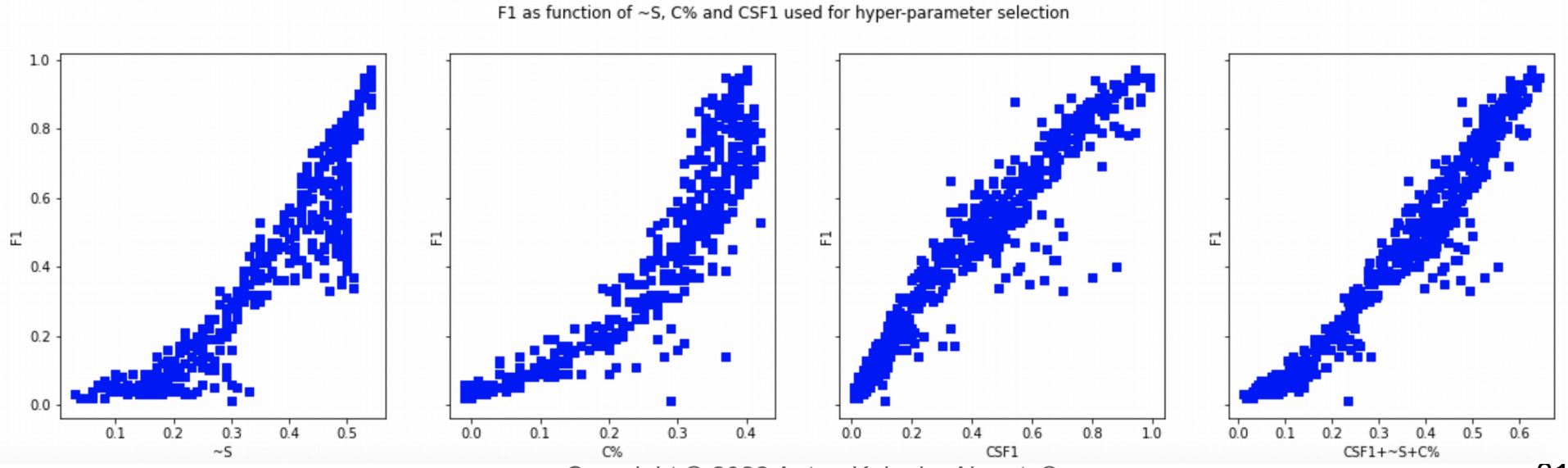
Lexicon-based Tokenization - greedy/beam search on word length (optimal) or frequency

Классификация знаков пунктуации, гласных и согласных в многоязычных текстах (по “свободам перехода”)

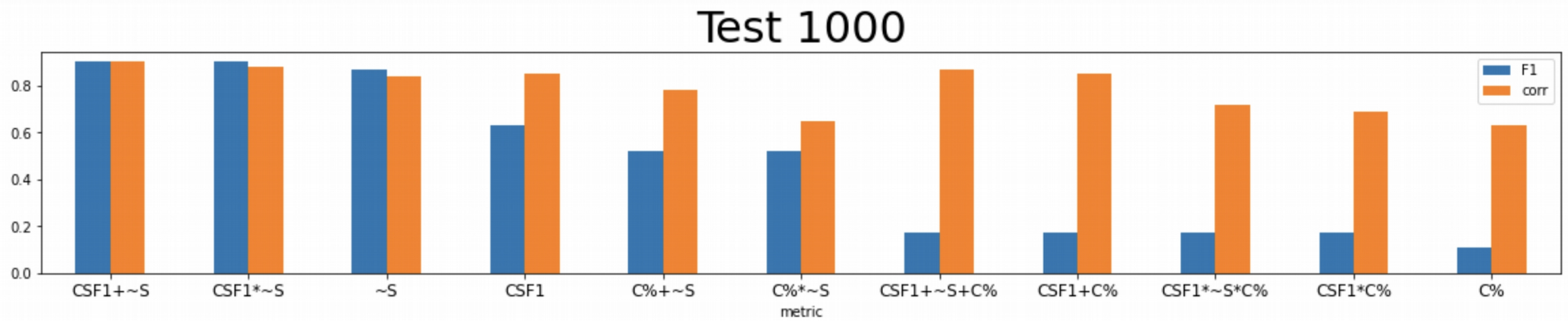


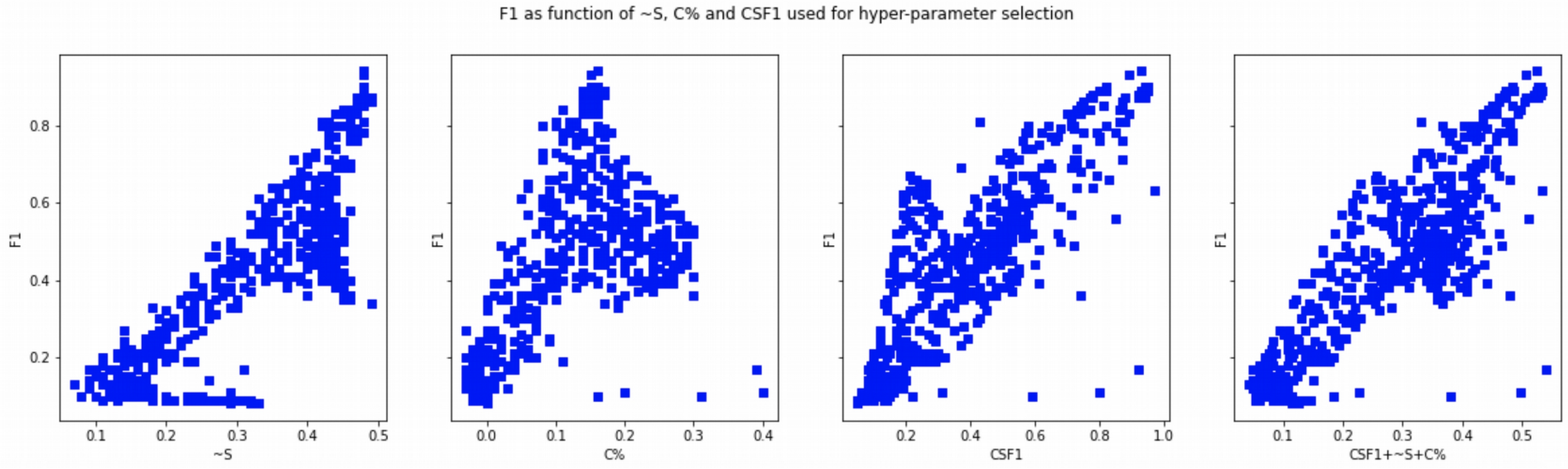
Самонастройка гиперпараметоров – Английский (отклонение TF)



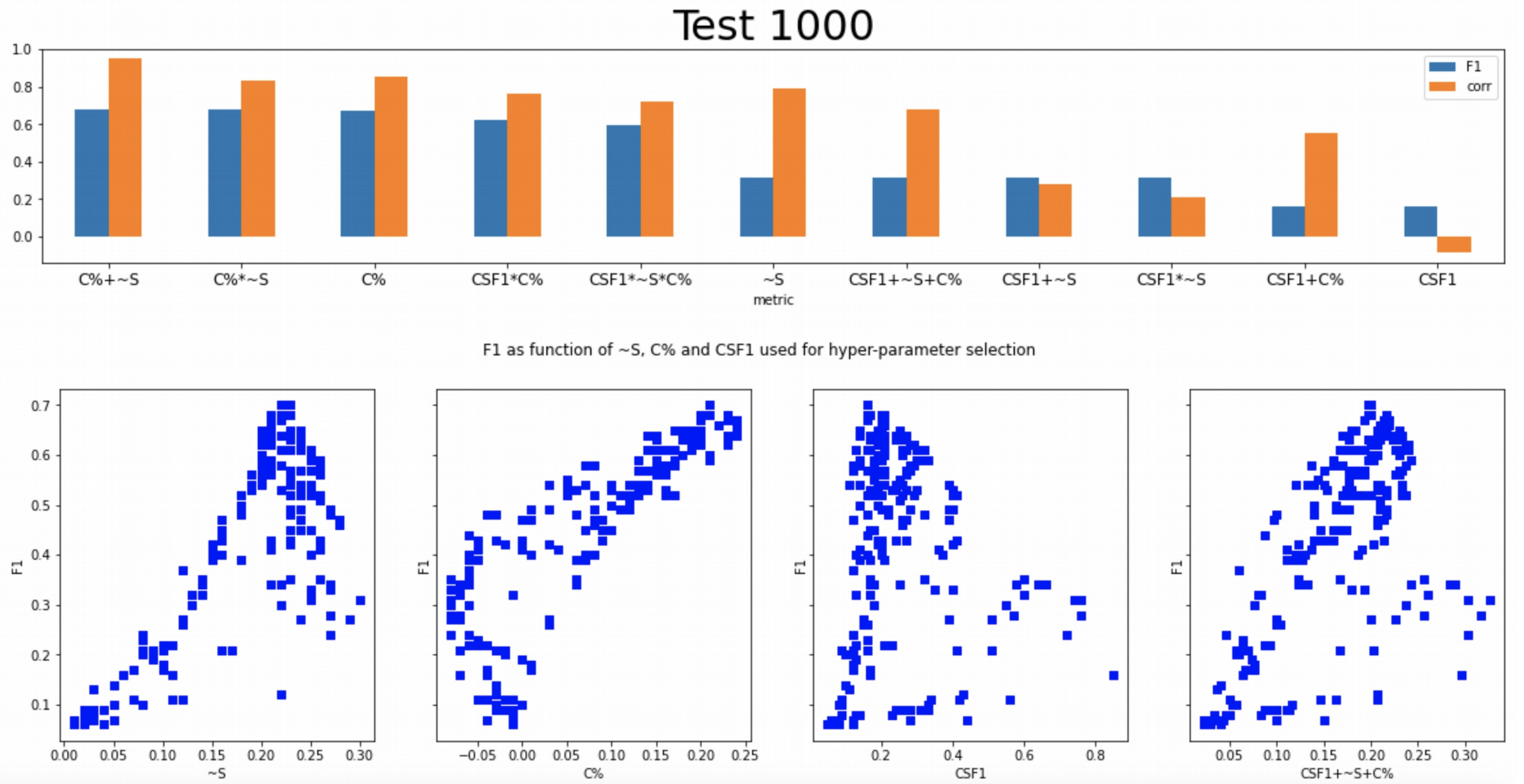


Самонастройка гиперпараметоров – Русский (отклонение TF)

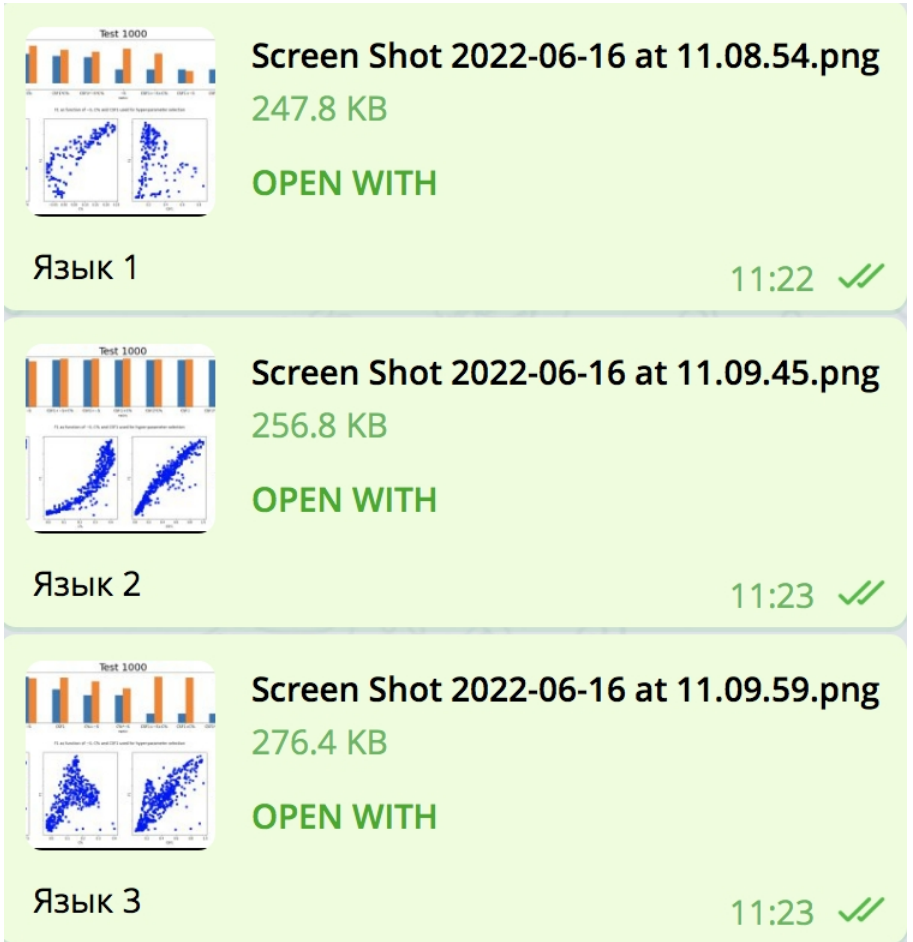


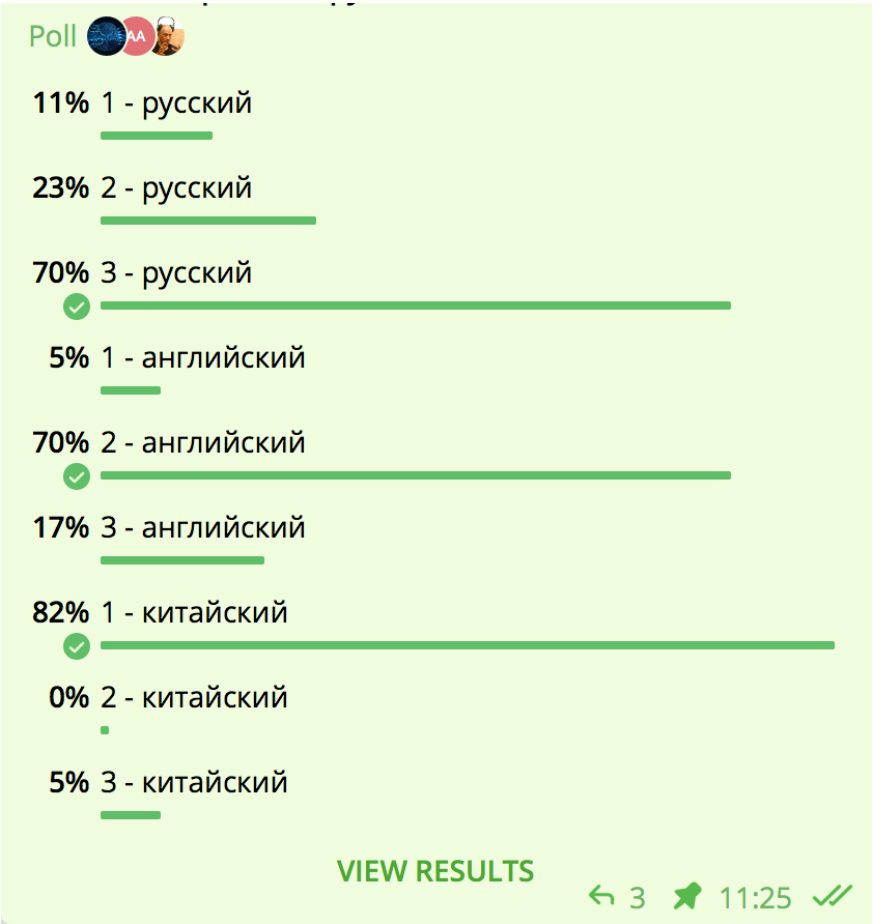


Самонастройка гиперпараметоров – Китайский (“пики” TF)



Вы все правильно угадали!





https://t.me/agibots/2695

Выводы?

“Интуитивная” (человекоподобная) кластеризация – возможна!
По крайней мере, для временных рядов с мультимодальными распредлениями.

Высокоточная сегментация и выявление лексиконов и знаков пунктуации без учителя для языков типа русского и английского – возможна!
С языками типа китайского – надо разбираться.

Посмотрим, как это все будет работать для обучения с подкреплением, прогнозирования временных рядов и построения интерпретируемых полноценных языковых моделей...

https://github.com/aigents/pygents

Спасибо! Вопросы?

Антон Колонин

akolonin@aigents.com

https://github.com/aigents

Telegram: akolonin

Novosibirsk State University

*THE REAL SCIENCE

https://www.nsu.ru/

AIGENTS

https://aigents.com

AGI IN RUSSIAN

https://agirussia.org