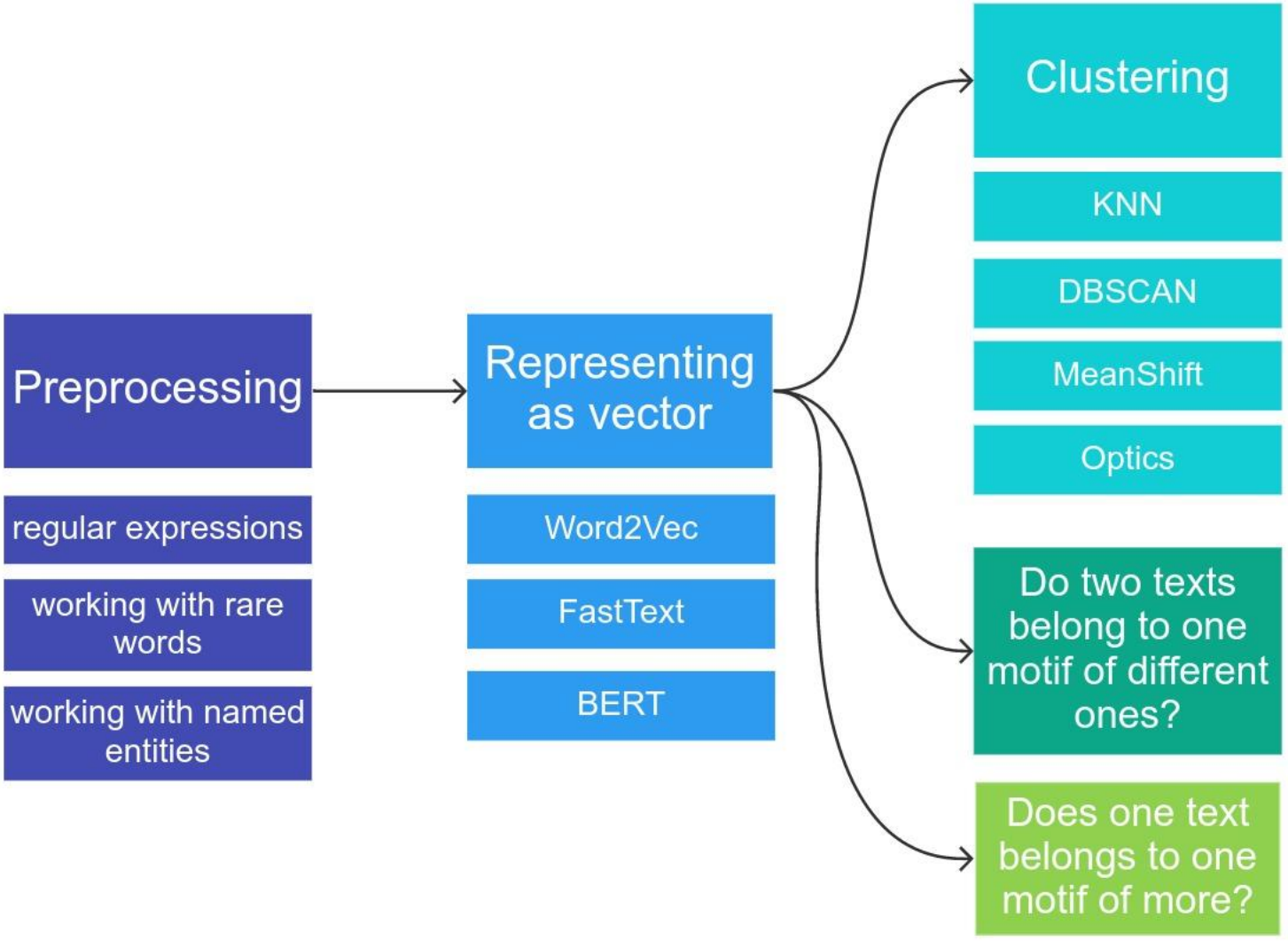
Text relation detection using hierarchical clustering techniques

M.Matveeva, NSU
2021

Dataset sample

```
final_df.sample(7)
```

	region	nation	text	code_name	motif	code_true	preprocessed
60318	тибет северовосток индии	тибетцы	(Амдо) [Кролик, M23, M62B, M120B, M152A];	Трикстер Заяц Кролик	См. мотивы, указанные в квадратных скобках. Вк...	M29G	[амдо, кролик]
33215	побережье плато	комокс	(чатлолтк) [две женщины Омак и Кьее...	Покинутый на острове	.19.20.22.37.-.47.49.50.59.61.74.	K1E	[чатлолтк, женщина, омак, кьеек, приезжать, ос...
65867	малайзия индонезия	саяки	(западный Саравак) [пландок (карликовый олень...	Посаженный в мешок	Персонажа кладут в мешок или сундук, запирают ...	M91c2	[западный, саравак, пландка, карликовый, олень...
57887	средняя европа	белорусы	[«Бычок смоляной бочок»: к нему прилип...	Звери откупаются	Человек ловит нескольких диких животных и отпу...	M182a1	[бычок, смоляной, бочок, прилипать, животное, ...
9160	тайвань филиппины	моро	[небо было так низко, что когда солнце проход...	Низкое небо	Небо было рядом с землей, затем поднялось.	B77	[небо, низко, солнце, проходить, свой, путь, п...
21427	китай корея	намузи	[трое братьев обрабатывают землю, утром земля...	Скот возвращается в воду	Получив от сверхъестественных персонажей домаш...	H18B	[трое, брат, обрабатывать, земля, утро, земля,...
60443	арктика	медные	[Сова, M123, M123C];	Трикстер Сова	См. мотивы, указанные в квадратных скобках.	M29H	[сова]


Preprocessing
regular expressions
working with rare words
working with named entities
Representing as vector
Word2Vec
FastText
BERT
Clustering
KNN
DBSCAN
MeanShift
Optics
Do two texts belong to one motif of different ones?
Does one text belongs to one motif of more?

Preprocessing: regular expressions

```
array([' [отец юноши дает ему надеть кожу ворона; он пролетает на небо сквозь узкий проход, стены которого то сход
ятся, то расходятся; пихтовая хвоинка; люди-духи исчезают при свете солнца]: Barbeau 1961: 83-85; '],
      dtype=object)
```

Preprocessing: working with rare words

- excluding words which occur less than 5 times in all data
  - advantage: helps to get rid of incorrectly written words, names, etc
  - disadvantage: many noise words occur a lot in the data (‘informant’, ‘motif’)

Preprocessing: working with named entities

- many rare words in my dataset are named entities
- DeepPavlov's BERT-based NER was used to find them
- can be excluded (rougher approach) or replaced with unified entities

```
print(ner_model(['Наташа приехала в Москву и сразу пошла в Третьяковскую галерею'])[0])
print(ner_model(['Наташа приехала в Москву и сразу пошла в Третьяковскую галерею'])[1])
```

```
[['Наташа', 'приехала', 'в', 'Москву', 'и', 'сразу', 'пошла', 'в', 'Третьяковскую', 'галерею']]
[['B-PER', 'O', 'O', 'B-LOC', 'O', 'O', 'O', 'O', 'B-ORG', 'I-ORG']]
```

Clustering

- KNN
  - trained on bit and small labels (0.51 and 0.13 accuracy respectively)
- MeanShift
- DBSCAN & OPTICS
  - make clusters of small texts and almost the same texts (case when one text belongs to different labels)

Do two texts belong to one motif of different ones?

Idea: using cosine distance between two texts, predict whether they have the same motif or not
0.60 ROC AUC

	vectors_left	text_left	code_true_left	vectors_right	text_right	code_true_right	distance	target
2787127	[-0.8181673288345337, 0.10940517485141754, 0.0...	пошел к берегу со своими шестью братьями, те б...	K27	[-0.38197264075279236, 0.3214205503463745, -0....	из отчета А.Е. Грена, материалов умирающий вел...	K27	6.105389e-02	1
2995207	[-0.4122503995895386, 0.2664409577846527, -0.5...	царь	K27N	[-0.4122503995895386, 0.2664409577846527, -0.5...	хозяйка	K27N	4.440892e-16	1
482984	[0.16112986207008362, 0.22820784151554108, 0.8...	в песнях зимнего календарного цикла: «Месяц Со...	A4	[-0.023050647228956223, -0.5968122482299805, 0...	звезды жены Месяца (пол Солнца не известен); с...	A4	2.180092e-01	1
10529	[-0.5611329078674316, -0.08034060150384903, -0...	братья рубят дерево на небе; упав, оно задавил...	C24	[-0.4716351330280304, 0.5593582391738892, -0.4...	хан получил послание: мой жеребец заржал, годо...	M114G	2.087645e-01	0
45249	[0.25215262174606323, -0.539249062538147, -0.5...	на пути душ в загробном мире камень со ртом (б...	L32	[-0.7298957705497742, 0.8525152206420898, -0.2...	см. мотив брат и сестра приходят к людям без а...	F49	2.754632e-01	0

Does one text belongs to one motif of more?

Idea: understand by text vector whether it is associated with one motif or many

1. Take 10 closest texts by cosine distance (KDTree)
2. Mark then as ‘have the same main motif’ (big letter) of not
3. Train the model

0.71 ROC AUC

Thank you for your attention!