Тематическое моделирование русско-китайского медиа-дискурса о Второй мировой войне (на материале публицистического корпуса 2021-2025гг.)

Итоговый проект МЛ Иванченко Надежда

Цель проекта:

тематическое моделирование и сравнение русскоязычного и китайскоязычного медиа-дискурса о Второй мировой войне с применением эмбеддингов и кластеризации.

Этапы проекта:

- ➤ Предобработка корпусов и разметка метками (label=cn, label=rus)
- ▶ Генерация эмбеддингов
- ➤ Построение модели BERTopic
- > Выделение тем с высокой вероятностью
- Интерпретация ключевых слов и присвоение заголовков
- > Сравнительный анализ тематического наполнения

Корпус статей 2021-2025 года:

~3,000 китайских статей (People.cn, запрос «抗日» (Антияпонское сопротивление))
~3,000 русских статей (Iz.ru, запрос «Великая Отечественная война»)

- Тексты статей очищены от мусора и содержат только текст (на русском и китайском языках), но не лемматизированы и не очищены от стоп-слов.
- В таблицы, содержащие "Заголовок", "Дата", "Текст статьи" добавлены метки для различения корпусов ("cn" и "rus").
- Удалены ячейки без текстов, которые «собрались» в процессе парсинга.

labe	Текст статьи	Дата	Заголовок
ru	В честь Международного дня освобождения узнико	11.04.2023 00:00	МО запустило проект ко Дню освобождения узнико
ru	Представители общественного движения «Народный	08.05.2024 14:34	редставители «Народного фронта» поздравили ве
ru	Артиллерийские снаряды времен Великой Отечеств	14.05.2024 13:34	Восемь снарядов времен ВОВ нашли около деревни
Ct	2022年06月02日08:23 来源:解放军报 写在前面 在战火纷飞的年代里,我军诸	2022-06-02 03:23	一个文艺战士的炽热情怀
CI	12月8日,吉林省委党史研究室主任穆占一率队到池南区实地调研全国爱国主义教育示范基地长 白山老	2021-12-10 04:10	吉林省委党史研究室到老黑河遗址实地调研
CI	昔日,这里商贾云集,是四川通往云贵的东大路,有3座码头成为历史印记;如今,这里保留着近300	2024-08-23 02:42	近300座明清传统建筑"隐居"永宁河畔
ru	Снимок с боевиками Вооруженных сил Украины (ВС	20.12.2022 23:15	Замминистра обороны Украины опубликовала фото
ru	В России хотят распространить внеклассные «Раз	03.10.2024 19:47	Дошкольникам пора: «Разговоры о важном» могут
ru	В России 1 июня отмечается День Северного флот	31.05.2024 00:00	1 июня— День Северного флота в России: истори

Модель эмбеддингов: paraphrase-multilingual-MiniLM-L12-v2 (модель Sentence-BERT) https://huggingface.co/sentencetransformers/paraphrase-multilingual-MiniLM-L12-v2 -

- Предобученная
- Мультиязычная
- легкая и компактная версия BERT (трансформере MiniLM с 12 слоями и ок. 118 млн параметров)
- Использует WordPiece токенизацию (целые слов и подслова – для русского и иероглифы – для китайского)
- работает с эмбеддингами на уровне предложений

```
Topic Count
      -1
          2559
                              -1 на что не по
               0_времен_снаряд_обнаружили_мчс
                   1_победы_мая_великой_парад
           135
                              2 扫荡 添加 微信 扫一扫
      2
           126
                         3 花儿 如今 分享让更多人看到 同时
      3
           115
      4
           107
                  4 день праздник году россии
      5
                             5 五一口号 因此 添加 微信
            96
                 6 музея музей выставка музее
                            7 添加 微信 扫一扫 学习时报
            77
                           8 全文如下 学习时报 添加 微信
            74
                             Representation \
      [на, что, не, по, это, для, он, из, как, за]
1 [времен, снаряд, обнаружили, мчс, обезвредили,...
2 [победы, мая, великой, парад, площади, войне, ...
3 [扫荡,添加,微信,扫一扫,创建模范机关,时习之,系列重要讲话数据库,学习大国...
4 「花儿,如今,分享让更多人看到,同时,近年来,口袋公园,目前, амударьи...
5 [день, праздник, году, россии, праздника, флот...
    [五一口号,因此,添加,微信,扫一扫,指出,强调,两个确立,习近平,来源]
7 [музея, музей, выставка, музее, на, выставки, ...
```

- - 8 [添加, 微信, 扫一扫, 学习时报, 创建模范机关, 时习之, 系列重要讲话数据库, 学习...
 - 9 [全文如下,学习时报,添加,微信,扫一扫,来源,创建模范机关,时习之,系列重...

Representative Docs

- [Максимальный размер дальневосточной ипотеки п...
- [В Орловской области сотрудники группы спецраб...
- [На военном параде в честь 79-й годовщины побе...
- 来源:人民政协报 1942年8月,八路军冀... [吴小龙 2023年12月03日09:16
- 「柴关乡王硇村风光。 在柴关乡石门沟村,街巷地面全由红石板铺成,石板街道随地势起伏,陡坡处垒...
- [Ежегодно 23 февраля по всей России отмечается...
- 来源: 光明日报 手机看新闻 习近平总... 「柳建辉 2024年11月05日15:29
- [Приблизить историю к сегодняшнему дню, буквал...
- [2022年07月10日09:09 来源:解放军报 陈列在甘肃省礼县龙池湾战役纪念馆里...
- 「辽沈战役纪念馆讲解员为参观学生讲解。\n辽沈战役纪念馆供图 中央红军长征出发纪念馆讲解员钟...

Кастомизация, шаг 2:

hdbscan_model = HDBSCAN(min_cluster_size=5, min_samples=10)

	Topic	Count	Name
0	-1	2609	-1_80_2022_25_10
1	0	145	0_infoorelru_76_nashbryanskru_12700
2	1	131	1_79_78_0043_125
3	2	113	2_23_1918_1917_1995
4	3	108	3_193810_19407_1940_202308240826
5	4	107	4_1125_20208_194019426014_3002021
6	5	100	5_1935_20221027_75_1954219569
7	6	78	6_3d_1812_travel_live
8	7	75	7_10005_1941925350055_8080_117
9	8	74	8_88140_7814020060_19712005_2040
10	9	69	9_wink_netflix_dc_the

```
Representation \
     [80, 2022, 25, 10, 70, 2023, 2024, 20, 15, 2025]
   [infoorelru, 76, nashbryanskru, 12700, krasnod...
   [79, 78, 0043, 125, 2200, 34, 1000, 1945, 400,...
   [23, 1918, 1917, 1995, xix, 1980, 1790, 1812, ...
   [193810, 19407, 1940, 202308240826, 1942210194...
   [1125, 20208, 194019426014, 3002021, 202330546...
    [1935, 20221027, 75, 1954219569, 19227194528, ...
   [3d, 1812, travel, live, 1883, 12, abrams, 187...
   [10005, 1941925350055, 8080, 117, 77, 91818, 1...
   [88140, 7814020060, 19712005, 2040, 201570, 19...
10 [wink, netflix, dc, the, 1929, marvel, star, m...
11 [mbbarklay, 1944, 1945, rm, gunna, 1947, 1942,...
12 [1967, 2014, 11917, krasnoyarskmedia, 1874, ch...
13 [0043, 2025, 80, 79, 1945, twitter, 89, 19, 70...
14 [930, 110110, 4582022, 200914, dna, 77, 93, 77...
15 [himars, baby, sea, fpv, 132, 155, antpq50, st...
16 [200660928, 192619292556741988, 530009000515, ...
17 [1418, 22, 83, 82, 0400, 400, 22061941, 1996, ...
18 [2015, ua, telegram, facebook, meta, lrt, 28, ...
19 [8867, 7519496, 1793, 372212842, 1953724742202...
```

Шаг 3 -BERTopic с кастомной настройкой:

- ➤ HDBSCAN для кластеризации ((min_cluster_size=7, min_samples=3),
- CountVectorizer с n-граммами (1— 3),
- индивидуальные списки стопслов для каждого языка (хотя, их можно было бы и дополнить)
- ▶ Включена вероятностная оценка принадлежности к теме (calculate_probabilities=True)

custom_stopwords = [

• Русские стоп-слова:

"и", "в", "во", "не", "что", "он", "на", "я", "с", "со", "как", "а", "то", "все", "она", "так", "его", "но", "да", "ты", "к", "у", "же", "вы", "за", "бы", "по", "ее", "мне", "было", "вот", "от", "меня", "еще", "нет", "о", "из", "ему", "теперь", "когда", "даже", "ну", "вдруг", "ли", "если", "уже", "или", "ни", "быть", "был", "него", "до", "вас", "нибудь", "опять", "уж", "вам", "ведь", "там", "потом", "себя", "ничего", "ей", "может", "они", "тут", "где", "есть", "надо", "ней", "для", "мы", "тебя", "их", "чем", "была", "сам", "чтоб", "без", "будто", "чего", "раз", "тоже", "себе", "под", "будет", "ж", "тогда", "кто", "этот", "этот", "это"

• Английские стоп-слова

"the", "and", "is", "in", "to", "of", "for", "on", "with", "as", "by", "at", "an", "be", "this", "from", "telegram", "facebook", "meta", "Sea Baby", "marvel", «wink", "netflix", "iii",

• Китайские стоп-слова

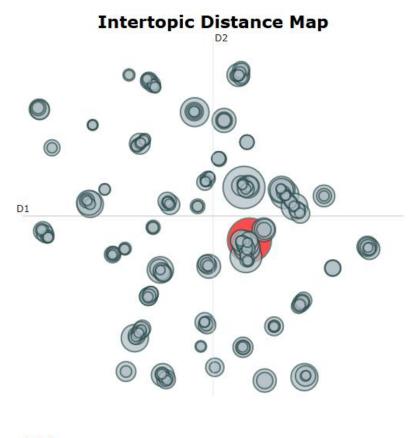
"的", "了", "在", "是", "我", "有", "和", "就", "不", "人", "都", "一个", "也", "很", "到", "说", "要", "你", "会", "着", "没有", "看", "好", "自己", "这", "那", "我们", "他", "她", "它", "被", "上", "中", "下", "对", "吧", "扫一扫", "添加", "微信", "分享让更多人看到", "打造", "指出", "全文如下", "时习之",

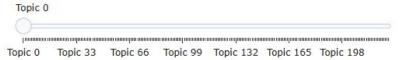
Результаты шага 3

	Topic	Count	Name \		
0	-1	2082	-1_россии_года_отечественной_великой		
1	0	144	0_времен_времен великой отечественной_времен в		
2	1	122	1_花儿_如今_同时_近年来		
3	2	88	2_中央和国家机关 创建模范机关_创建模范机关_习近平 系列重要讲话数据库 中央文件	+_学习)路
4	3	78	3_памятник_памятников_памятника_снос		
5	4	72	4_победы_мая_президент_победы великой отечеств		
6	5	64	5_随后_赓续红色血脉_胜利渡长江_缅怀革命先烈		
7	6	64	6_могила_отечественной войны_великой отечестве		
8	7	59	7_例如_推动红色旅游高质量发展_赓续红色血脉_同时		
9	8	58	8_памяти_свеча_свеча памяти_акции		
10	9	55	9_扫荡_八办_中央和国家机关 创建模范机关_创建模范机关		Rep
11	10	54	10_день_праздник_году_россии	0	года, отечественной, великой,
12	11	54	11_поздравили_военнослужащие_ветерана_отечеств		[времен, времен великой отечественной,
13	12	52	12_суд_геноцидом_области_территории	2	[花儿,如今,同时,近年来,目前,口袋公园
14	13	50	13_来源 学习时报_学习时报_中央和国家机关 创建模范机关_创建模范机		[中央和国家机关 创建模范机关, 创建模范机关
15 16	14 15	48 46	14_游览地_建议时长_一日游_建议时长 一日游 15_пенсии_пенсию_выплаты_рублей	4	[памятник, памятников, памятника, снос
17	16	46	16_путин_россии_победы_президент	5	[победы, мая, президент, победы велико
18	17	44	17_морской_флота_кораблей_вмф	6	[随后,赓续红色血脉,胜利渡长江,缅怀革命
19	18	39	18 войска известий дней фронта	7	[могила, отечественной войны, великой
	10		zo_boneita_nbbee1mi_Anen_appointa	8	[例如,推动红色旅游高质量发展,赓续红色血
					[памяти, свеча, свеча памяти, акции, с
					[扫荡,八办,中央和国家机关 创建模范机关,
					[день, праздник, году, россии, отмечаю
					[поздравили, военнослужащие, ветерана,
					[суд, геноцидом, области, территории,
					[来源 学习时报,学习时报,中央和国家机关
					[游览地,建议时长,一日游,建议时长 一日》
					[пенсии, пенсию, выплаты, рублей, разм
					[путин, россии, победы, президент, поз
					[морской, флота, кораблей, вмф, корабл
				19	[войска, известий, дней, фронта, рамка

Распределение максимальных вероятностей тем 2000 Кол-во документов 1500 1000 500 0.0 0.2 0.6 8.0 1.0 0.4 Макс. вероятность

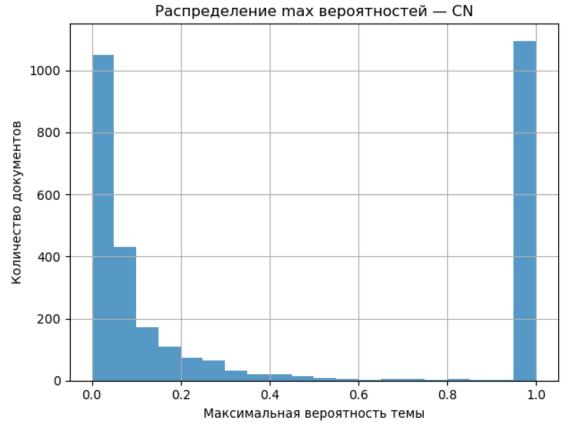
Отфильтруем слабоуверенные тексты (threshold = 0.3) и посмотрим на визуализацию

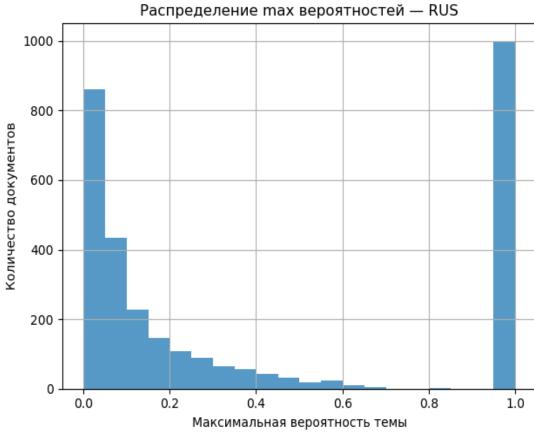




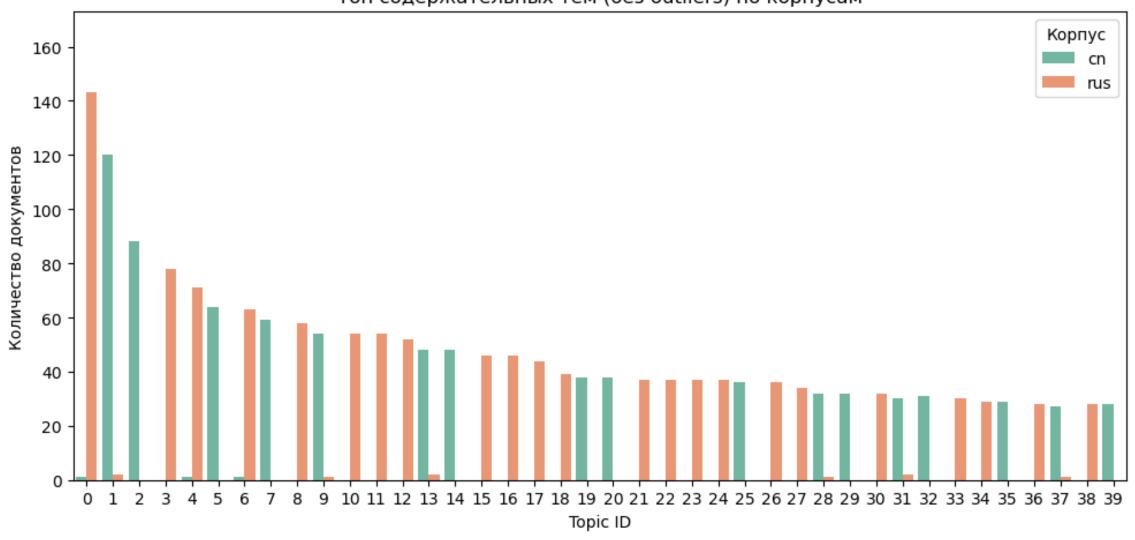
Шаг 4. Разделим корпуса и обучим две модели отдельно на русском и отдельно на китайском языках

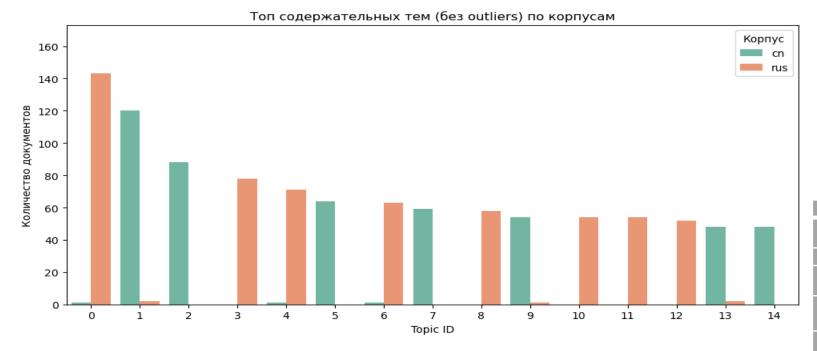
```
topic_model_cn = BERTopic(
   hdbscan_model=hdbscan_model,
   vectorizer_model=vectorizer_model,
   language="multilingual",
   calculate_probabilities=True
)
topics_cn, _ = topic_model_cn.fit_transform(docs_cn, model.encode(docs_cn, batch_size=32))
```





Топ содержательных тем (без outliers) по корпусам





	Интерпретация темы (русский)			
•				
0	боеприпасы времен ВОВ			
1	выставки и экспозиции о войне			
2	Могилы солдат ВОВ			
3	Демонтаж памятников ВОВ			
4	Военные преступления, геноцид			
5	Парад Победы			
6	Празднование победы			
7	Пенсии и выплаты			
8	Рассказы ветеранов войны			
9	Акция "бессмертный полк"			
10	Военно-морской флот			
11	Красная армия, 1943 год			
12	Поздравление лидера страны			
13	Театр			
14	Отношения с соседями (Беларусь, Узбекистан)			

	Интерпретация темы (китайский)
0	выступления Си Цзиньпина и центральные документы
1	Центральные и государственные органы
2	Обучение и изучение цикла речей Си Цзиньпина
3	Праздник Цинмин и почтение памяти революционных мучеников
4	выставки посвященные героям в войне против Японии (гора Ланья, компания «Люлаочжуан»)
5	Фильмы о войне и военных героях (Постановка/фильм "Большой нож", "Сто рек текут на восток, к морю")
6	Места для посещения, туризм
7	культурные выставки (народных промыслов и картин)
8	Тематические мероприятия (Путешествие во Внутреннюю Монголию)
9	Центральные и государственные органы; изучение речей Си Цзиньпина
10	героический дух борьбы, боевой марш
11	представители и представительства; муниципалитет, организации
12	военная поэзия
13	скорбь и поминальные мероприятия
14	военная фотография

Выделенные cnмоделью темы

- 扫荡,中央和国家机关创建模范机关,创建模范机关,习近平系列重要讲话数据库中央文件,学习路上习近平 库中央文件,学习大国
- 中央和国家机关 创建模范机关, 创建模范机关, 习近平 系列重要讲话数据库, 学习路上 习近平 系列重要讲话数 据库, 系列重要讲话数据库, 系列重要讲话数据库 中央文件, 学习路上 习近平, 习近平 系列重要讲话数据库 中央 Центральные и государственные 文件, 学习大国, 学习路上
- 来源 学习时报, 学习时报, 创建模范机关, 中央和国家机关 创建模范机关, 习近平 系列重要讲话数据库, 系列重 要讲话数据库, 学习路上 习近平 系列重要讲话数据库, 习近平 系列重要讲话数据库 中央文件. 学习路上 习近平. Обучение и изучение цикла речей Си 系列重要讲话数据库 中央文件
- 随后, 赓续红色血脉, 缅怀革命先烈, 敬献鲜花, 钢八连, 9月3日, 献花曲, 传承红色基因, 清明节的铭记, 清明祭英

工人之路,狼牙山五壮士,狼牙山,刘老庄连,本报记者,烽火印记北京抗日战争主题片区特展,烽火印记,北京抗 日战争主题片区特展,例如,石洪宇

大刀记, 红湾, 百川东到海, 黄宗德说, 十年来, 内容简介, 临郯青年救国团团歌, 戏中戏, 热血当歌, 中央和国家机 关 创建模范机关

建议时长,游览地,建议时长一日游,一日游,榆次区,游览地榆次区,游览地榆次区建议时长,榆次区建议时长 00,建议时长二日游

近年来,如今,昆嵛,梳乡,之称,山东手造,文化,连城县委组织部供图,阳泉解放建市,记载

- 彻党的二十大精神 主题展, 学习宣传贯彻党的二十大精神, 主题展
- 刘志丹, 中央和国家机关 创建模范机关, 创建模范机关, 学习路上 习近平 系列重要讲话数据 讲话数据库, 学习大国
- 机关 创建模范机关
- 属代表 中央党政军群有关部门和北京市负责同志, 在京参加过抗日战争的老战十及抗战将领亲属代表 抗战烈 士遗属代表 中央党政军群有关部门和北京市负责同志

12

14

七律,讲话,沁园春,五律,四言诗,比如,长征,诗言志,在延安文艺座谈会上的讲话,写道

行祭扫活动 缅怀先辈 寄托哀思, 各地举行祭扫活动, 缅怀先辈 寄托哀思, 清明节将至, 缅怀先辈, 寄托哀思, 4月2 \Box

凯尔说, 受访者供图, 历史像一面镜子不正视它的人注定失败, 这些照片属于博物馆, 这些照片 被公之于众, 历史像一面镜子, 不正视它的人注定失败, 必须被公之于众, 表达我对这片土地的敬意, 埃文

выступления Си Цзиньпина и центральные документы

органы

Цзиньпина

Праздник Цинмин и почтение памяти

революционных меченников выставки посвященные героям в войне против Японии (гора Ланья, компания «Люлаочжуан»...) Фильмы о войне и военных героях (Постановка/фильм "Большой нож", "Сто рек текут на восток, к морю"...)

Места для посещения, туризм культурные выставки (народных промыслов и картин)

Тематические мероприятия -大精神 主题展, 学习盲传贯 (Путешествие во Внутреннюю Монголию)

> ,习近平 系列重要 Центральные и государственные 习近平 系列重要 органы; изучение речей Си Цзинтпина героический дух борьбы, боевой марш

> > муниципалитет, организации

военная поэзия

скорбь и поминальные мероприятия

военная фотография

Выделенные rus-моделью темы

0	времен, времен великой отечественной, времен великой, мчс, обнаружили, снаряд, боеприпасов, великой отечественной войны, отечественной войны, войны	боеприпасы времен ВОВ
1	музея, музей, выставка, музее, выставки, победы, можно, москвы, истории, экспозиции	выставки и экспозиции о войне
2	могила, саур, саур могила, году, великой, великой отечественной, отечественной, войны, отечественной войны, великой отечественной войны	Могилы солдат ВОВ
3	памятник, памятников, памятника, демонтировали, снос, демонтаж, советским, советских, власти, об	Демонтаж памятников ВОВ
4	суд, геноцидом, области, территории, немецко, признал, геноцида, тыс, суд признал, преступления	Военные преступления, геноцид
5	победы, мая, площади, красной площади, парада, парад, красной, победы великой отечественной, великой отечественной войне, победы великой	Парад Победы
6	победы, мая, президент, победы великой отечественной, победы великой, великой отечественной войне, отечественной войне, празднование, москве, войне	Празднование победы
7	пенсии, пенсию, выплаты, рублей, размер, пенсия, пенсий, страховой, старости, выплат	Пенсии и выплаты
8	ветеран, войны, лет, году, мая, года, отечественной, после, рассказал, время	Рассказы ветеранов войны
9	бессмертный полк, бессмертный, полк, мая, акции, акция, победы, шествие, акция бессмертный, акция бессмертный полк	Акция "бессмертный полк"
10	морской, флота, кораблей, вмф, корабль, адмирал, корабли, морской пехоты, пехоты, боевых	Военно-морской флот
11	армии, немцы, армия, года, красной армии, 1943, красная армия, красная, красной, 1943 года	Красная армия, 1943 год
12	путин, россии, президент, победы, поздравил, государства, лидер, отечественной, президент россии, великой отечественной	Поздравление лидера страны
13	году, артист, лет, озеров, песни, певец, очень, после, театре, актриса	Театр
14	путин, президент, рф, лукашенко, россии, стран, двух, владимир путин, мая, узбекистана	Отношения с соседями (Беларусь, Узбекистан)

谢谢关照