Новый датасет для решения задачи фильтрации чужого текста в социальных медиа

Ивойлова А.М. (a.m.ivoylova@gmail.com)

РГГУ, Москва, Россия

Раскин И.И. (raskin.ii@phystech.edu)

ABBYY Lab, МФТИ, Москва, Россия

Селегей Д.В. (daniil.selegey@abbyy.com)

ABBYY Lab, МФТИ, Москва, Россия

Статья основывается на исследовательской работе, которую проводят студенты МФТИ и РГГУ в рамках проекта создания Генерального Интернет-Корпуса Русского Языка (ГИКРЯ). В настоящее время одна из актуальных задач в области разработки мегакорпусов по технологии Web as Corpora — это создание репрезентативного корпуса, очищенного от нерелевантных текстов (спама, цитирования, ссылок и т. п.)

Главная цель данного проекта — создание обучающего датасета классифицированных по категориям «грязных» текстов из социальных сетей для дальнейшего анализа и работы по очистке подкорпуса ГИКРЯ, включающего тексты из социальных сетей, от неавторских текстов. В статье описываются принципы работы и полученные результаты.

A New Dataset to Solve the Task of Text Filtration in Social Networks-Based Corpora

Ivoylova A. (a.m.ivoylova@gmail.com)

RSUH, Moscow, Russia

Raskin I. (raskin.ii@phystech.edu)

ABBYY Lab, MIPT, Moscow, Russia

Selegey D. (daniil.selegey@abbyy.com)

ABBYY Lab, MIPT, Moscow, Russia

This paper is based on research carried out by the students of MIPT and RSUH within the project on the General Internet Corpus of Russian (GICR). Currently, one of major tasks in web corpus building is to create a corpus that would be as clean as possible (which means containing no spam, duplicate texts, links and so on).

The main goal of our project is to create a learning dataset which consists of category-classified text types that need to be removed from a clean corpus. This dataset is to be subjected to further analysis and can be used in cleaning the GICR subcorpus based on texts from social networks. In this paper we describe the principles of our research and present our current results.

1. Введение

С развитием технологий корпусная лингвистика в последние годы сделалась одним из важнейших направлений в современных исследованиях; одна из основных причин такой популярности — это большие объемы данных и возможность использовать достаточно репрезентативный источник (Пиперски, 2013). Такими источниками с середины XX века служили традиционные закрытые корпуса, а в последнее время благодаря совместной работе лингвистов и инженеров появляются так называемые мегакорпуса (Беликов и др., 2014), или сверхбольшие корпуса (Benko, Zakharov, 2016): это корпуса текстов, составленные при помощи технологии Web as Corpus (WaC) и содержащие больше миллиарда слов. Такие корпуса теоретически могут стать наиболее репрезентативными и представлять собой язык целиком, а не только какую-то его часть (обычно жанрово предопределенную). Генеральный Интернет-Корпус Русского Языка (ГИКРЯ) (Belikov et al., 2013), работе с которым посвящена данная статья, является одним из таких мегакорпусов. Характерная особенность ГИКРЯ состоит в том, в корпусе тексты разделены по источникам (так называемым сегментам интернета) (Беликов и др., 2012), что немаловажно, т. к. позволяет в исследованиях учитывать и то, откуда были взяты лингвистические данные. Так, социальные сети часто служат базой для социолингвистических исследований разного рода, и важнейшим преимуществом подобного источника данных является возможность использовать разные дифференциальные признаки экстралингвистического характера: возраст, гендер, регион, социальное положение и др.

Однако основная проблема всех корпусов, составленных по технологии WaC, в частности, и корпуса на основе текстов социальных сетей — это загрязненность текстов: автоматически собранные документы обычно содержат большое количество различного спама и мусора, не имеющего лингвистической значимости. О влиянии спама на результаты выдачи писали многие исследователи (н-р, этими вопросами занимаются V. Baisa, V. Suchomel (Baisa, Suchomel, 2012; Suchomel, 2017), а также A. Kilgariff (Kilgariff, Suchomel, 2013)), а в случае корпуса текстов социальных сетей задача осложняется обилием репостов, ссылок и прочего шума. Об одной из самых серьезных проблем мегакорпусов — проблеме дублирования и скрытого цитирования — еще в 2012 году писал В. И. Беликов (Беликов и др., 2012). Очевидно, что все эти явления искажают не только показатели частотности токенов, но и результаты различных социолингвистических исследований, связанных с изучением социо- или диалектов.

Таким образом, создание «чистого» корпуса — одна из главных на сегодня задач в корпусной лингвистике. Небольшой корпус можно обработать вручную, однако мегакорпус в несколько миллиардов слов может быть обработан только автоматически. Эта проблема до сих пор остро стоит перед современными исследователями (Jakubíček et al., 2020); одна из последних разработок — очищенный от лишних текстов и полностью размеченный англоязычный корпус Джорджтаунского университета AMALGUM (Gessler et al., 2020) — содержит в себе только 4 миллиона слов. Работа над созданием подобного «чистого» русскоязычного корпуса на основе текстов социальных сетей (в частности, социальной сети «ВКонтакте») ведется сейчас в рамках проекта ГИКРЯ совместными усилиями студентов МФТИ и РГГУ.

2. Путь к решению задачи

Как создать «чистый» корпус на основе текстов социальных сетей? Один из вариантов — собрать датасет из классифицированных образцов «шума», который в дальнейшем можно было бы использовать как для теоретического анализа, так и для машинного обучения. Простые способы решения этой задачи, предложенные предыдущими исследователями, оказались не слишком успешными; попытка собрать датасет из дублей (на основании предположения, что чистые тексты стремятся к уникальности) значимых результатов не принесла, а при использовании сервиса Яндекс.Толока получился очень недостоверный корпус.

В проекте, которым занимаются авторы данной статьи, было решено провести параллельное ручное и компьютерное исследование и сравнить результаты; таким образом, в проекте задействованы как программисты, так и квалифицированные разметчики-лингвисты (студенты МФТИ и РГГУ). Компьютерное исследование состоит из следующих шагов: в первую очередь проводится дедупликация текстов, в результате которой убираются не только полные дубли, но и так называемые near-duplicates (нечеткие дубликаты) (Benko, 2019): тексты с одинаковыми сигнатурами, свидетельствующими о шаблонном характере их производства (например, имеющие одинаковые начальные и/или конечные цепочки). На втором этапе проводится фильтрация с помощью дополнительных сигнатурных правил, используемых при очистке ГИКРЯ, и с использованием библиотеки fastText (https://fasttext.cc/).

Что касается проводимого вручную исследования, то здесь основная задача — в результате первичного анализа получить классификатор типов текстов и в дальнейшем разметить как можно большее их число (желательно до 100 тыс. текстов).

3. Принципы разметки и проверка согласия аннотаторов

В результате первичного анализа имеющихся документов была разработана система меток, представленная в табл. 1; примеры текстов для основных типов приведены в приложении 1.

Табл. 1. Система меток

author	тексты, написанные человеком, опубликовавшим пост
citation	цитирование (чужие высказывания и т. п.)
partial_author	авторский комментарий к тексту иной категории
article	статья; объемный текст, поддающийся тематической классификации
poem	стихотворение, в т. ч. написанное в строчку
fiction	объемные художественные тексты
news	новостные публикации
advertising	тексты рекламного характера, объявления, призывы
autogen	автоматически сгенерированные тексты
link_header	заголовок ссылки
foreign	текст полностью на любом языке, кроме русского
other	тексты, состоящие только из символов, ссылки, хэштеги или геотеги; прочий мусор

Как пишет С. А. Шаров в статье о функциональных жанрах (Sharoff, 2018), текст — явление многомерное, и бывает трудно приписать одному тексту ровно одну метку; все

то же справедливо и для системы меток в данном проекте. Первоначальная система меток жестко требовала ставить только один тег, что вызывало серьезные споры между разметчиками и низкий процент согласия; впоследствии было решено использовать гибридные метки с возможностью выбирать до трех меток и даже более (однако практика показывает, что больше двух меток ставится только в сложных и редких случаях). Это несколько осложнило проверку согласия разметчиков, т. к. в таком случае выбранные метки могут совпадать только частично, однако в целом процент расхождений снизился.

Имеющаяся выборка, полученная кроулингом «ВКонтакте» и состоящая в основном из текстов 2014–2016 гг., в процессе работы делится на датасеты размером в 1000 текстов каждый, после чего датасеты размечаются парами аннотаторов. Когда двойная разметка заканчивается, аннотаторы должны между собой обсудить тексты, размеченные неодинаково, и снять разногласия. Первые (тестовые) датасеты на данный момент размечены пятью или шестью аннотаторами; каждый раз проводилось обсуждение.

Как видно на рис. 1, процент по-разному размеченных текстов постепенно снижается (неожиданное единодушие разметчиков на втором датасете было вызвано тем, что аннотация проводилась по сути совместно; также первые пять датасетов были меньшего размера). Скачки после датасета №10 объясняются включением в работу новых разметчиков.

Рис. 1. Количество расхождений в разметке; по вертикали — процент расхождений (от объема датасета), по горизонтали — номер датасета

4. Текущее состояние проекта и первые полученные результаты

В настоящее время в проекте участвует девять разметчиков, в частности, студенты бакалавриата и магистратуры РГГУ (направления подготовки соответственно «Фундаментальная и прикладная лингвистика» и «Фундаментальная и компьютерная лингвистика»). Было размечено в сумме 18 848 текстов хотя бы двумя разметчиками; это количество продолжает увеличиваться. Диаграмма на рис. 2 иллюстрирует количество текстов разных категорий после ручной аннотации.

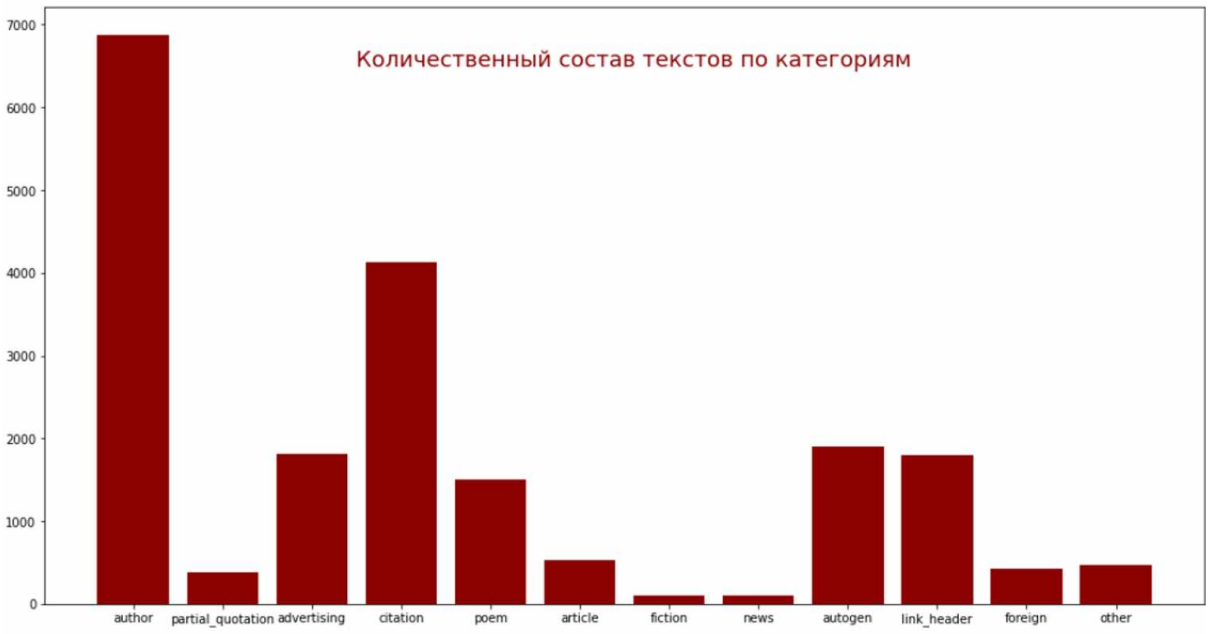

Рис. 2. Количественный состав текстов по категориям

Наибольшие затруднения у разметчиков вызывают категории citation и link_header; как правило, аннотатор не может достоверно присвоить метку тексту подобного рода, не проверив его при помощи поисковой системы (хотя цитаты и заголовки часто угадываются интуитивно, иногда интуиция приводит к ошибке). Также определенные разногласия вызывает категория advertising, возможно, из-за не слишком удачного названия (С. А. Шаров высказал предположение, что более подходящим было бы название promotion).

На имеющихся уже аннотированных данных были проведены последовательно процедура дедупликации и процедура фильтрации с помощью fastText.

В табл. 2 представлена текущая статистика по датасету с проведенной дедупликацией и фильтрацией на имеющихся на данный момент аннотированных вручную текстах. В первой колонке таблицы размещается название категории, где author — тексты с единственным тегом author, author+ — тексты, имеющие среди прочих этот тег, partial_author — тексты с данным тегом, trash — все остальные тексты, не имеющие тега author или partial_author; далее тексты остальных категорий. В двух следующих колонках указаны количество и доля текстов, имевшиеся в изначальном датасете, а затем — процент текстов, их количество и доля, оставленные соответственно после дедупликации и фильтрации. Надо отметить, что в сумме доля текстов по категориям не дает 100%, т. к. тексты могли попадать в несколько категорий одновременно.

Табл. 2. Результаты проведенной дедупликации и фильтрации на новом датасете

Название категории	Изнач. кол-во текстов	Изнач. доля текстов, %	%, оставляемый дедупликацией	Кол-во текстов после дедупликации	Доля текстов после дедупликации, %	%, оставляемый фильтрацией	Кол-во текстов после фильтрации	Доля текстов после фильтрации, %
ВСЕ	**18848**	**100**	**35,5**	**6689**	**100**	**84,2**	**5634**	**100**

author	**6420**	**34,1**	**75,9**	**4873**	**72,9**	**92,6**	**4511**	**80,1**
author+	**454**	**2,4**	**58,4**	**265**	**4,0**	**70,6**	**187**	**3,3**
partial_author	**363**	**1,9**	**38,8**	**141**	**2,1**	**74,5**	**105**	**1,9**
trash	**11611**	**61,6**	**12,1**	**1410**	**21,1**	**58,9**	**831**	**14,7**
advertising	1586	8,4	22,4	355	5,3	69,0	245	4,3
citation	4087	21,7	7,8	319	4,8	78,1	249	4,4
poem	1468	7,8	11,4	167	2,5	10,2	17	0,3
article	501	2,7	20,2	101	1,5	54,5	55	1,0
fiction	107	0,6	15,0	16	0,2	87,5	14	0,2
news	108	0,6	39,8	43	0,6	69,8	30	0,5
autogen	1894	10,0	6,8	129	1,9	42,6	55	1,0
link_header	1772	9,4	11,7	207	3,1	95,7	198	3,5
foreign	427	2,3	36,1	154	2,3	6,5	10	0,2
other	388	2,1	8,0	31	0,5	71,0	22	0,4

Как можно видеть, в исходном корпусе аннотаторы выделили 61,6% лишних документов (см. изначальную долю текстов). В результате автоматической обработки осталось только 14,7%; в основном это тексты, относящиеся к меткам, вызывавшим у аннотаторов наибольшие затруднения: citation, advertising и link_header.

Некоторые категории текстов, подлежащих отсеву, лучше удаляются дедупликацией, некоторые — фильтрацией; так, заголовки ссылок в основном убираются при помощи дедупликации, тогда как тексты на других языках устраняются фильтрацией.

Таким образом, убирается значительное количество «грязных» текстов, однако после процедуры дедупликации теряется и 24% текстов, размеченных аннотаторами вручную как авторские, а после фильтрации — еще 7,4%. Вероятнее всего, часть из них — это короткие часто повторяющиеся тексты (наиболее частотные из них — это «в точку!», «так и есть...» и подобные); возможно, сюда также попадают цитаты, не распознанные разметчиками, т. к. в случае ручной работы неизбежен некоторый процент ошибок. Так или иначе, этот набор текстов необходимо проанализировать.

5. Перспективы исследования

В процессе разработки системы меток для проекта было установлено, что качественная аннотация текстов не может быть осуществлена без отдельной метки partial_author (частичное цитирование), т. к. некоторое количество текстов в корпусе является авторским комментированием чужого текста. Порой отделить эти фрагменты почти не представляется возможным; во многих случаях аннотатор-человек видит авторский текст в окружении неавторского, но автоматическим способом отделить это не получается. Таким образом, естественное дальнейшее направление исследования — это вопрос об «авторских» переходах, т. е. решение задачи отделения авторского комментария от неавторского текста.

Литература

Baisa, Suchomel, 2012 – Baisa V, Suchomel V. (2012) Detecting Spam in Web Corpora. In Recent Advances in Slavonic Natural Language Processing. Proceedings to Sixth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, Karlova Studánka, Czech Republic, December 7–9, 2012.

Belikov et al., 2013 – Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S. (2013) Corpus as language: from scalability to register variation. In Dialogue, Russian International Conference on Computational Linguistics, Bekasovo.

Benko, 2019 – Benko V. (2019) Deduplication in Large Web Corpora. In Bański, Piotr/Barbaresi, Adrien/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Kupietz, Marc/Lüngen, Harald/Iliadi, Cadoline (Eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22 July 2019. – Mannheim: Leibniz-Institut für Deutsche Sprache, 2019. Pp. 17-21.

Benko, Zakharov, 2016 – Benko V., Zakharov V. P. (2016) Very Large Russian Corpora: New Opportunities and New Challenges. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016” Moscow, June 1–4, 2016.

Gessler et al., 2020 – Luke Gessler, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad, Amir Zeldes (2020) AMALGUM – A Free, Balanced, Multilayer English Web Corpus. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, 11–16 May 2020.

Jakubíček et al., 2020 – M. Jakubíček, V. Kovář, P. Rychlý, V. Suchomel. (2020). Current Challenges in Web Corpus Building. In Proceedings of the 12th Web as Corpus Workshop, Marseille, 11–16 May 2020.

Kilgariff, Suchomel, 2013 – Kilgariff A., Suchomel V. (2013) Web Spam. In Web as Corpus Workshop (WAC-8).

Sharoff, 2018 – Sharoff, Serge (2018) Functional Text Dimensions for the Annotation of Web Corpora. Corpora. 13. 65-95. 10.3366/cor.2018.0136.

Suchomel, 2017 – Suchomel V., Removing Spam from Web Corpora Through Supervised Learning Using FastText (2017) In Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/Stevenson, Mark/Sick, Theresa (Eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC-XI) guest section. Birmingham, 24 July 2017. - Mannheim: Institut für Deutsche Sprache, 2017. Pp. 56-60.

Беликов и др., 2012 – Беликов В., Селегей В. П., Шаров С. А.(2012) Пролегомены к проекту Генерального интернет-корпуса русского языка (ГИКРЯ). Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной конференции Диалог, Бекасово.

Беликов и др., 2014 – Беликов В., Копылов Н., Пиперски А., Селегей В., Шаров С. (2014) Дифференциальная корпусная статистика на основании неавтоматической

метатекстовой разметки. Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной конференции Диалог, Бекасово.

Пиперски, 2013 – Пиперски А.Ч. (2013) Генеральный интернет-корпус русского языка и понятие репрезентативности в корпусной лингвистике. Институт лингвистики ФГБОУ ВПО «Российский государственный гуманитарный университет», Москва, Россия.

Приложение 1

1. Примеры текстов по некоторым категориям используемой в проекте классификации типов

author

(1) У меня иммунитет моржа, и температура поднимается раз в 5 лет. Так что, когда это происходит, я сразу начинаю взвешивать за и против кремации.

(2) Рудик и ведущие CRM- щики страны на минуточку, приехавшие обучать одесское подразделение новым штукам...очень серйозные, я прям не знала как вести себя...))))

citation

(3) "Из двух друзей всегда один раб другого , хотя часто ни один из них в этом себе не признается " М.Ю.Лермонтов "Герой нашего времени"

(4) Молчи, пока ты не в состоянии сказать нечто такое, что полезнее твоего молчания. Архимед.
#вдушевесна#спасибочторядом#девочкитакиедевочки#любовь#весна#красота#вперед итольковперед#ипустьудачавсегдабудетсвами

partial_author (выделен неавторский текст)

(5) "**А прогнать меня ты уже не сумеешь. Беречь твой сон буду я**"... Вы сумасшедшие и восторге от постановки #МастериМаргарита в #МХТ . Так необычно, так свежо. Открыла для себя новый талантище - Михаил #Трухин. Не ожидала, что он может так играть. И #Белый... Безусловно..... #Булгаков #Рукописинегорят #Назаров #ПонтийПилат

(6) **Чем у женщины хуже с личной жизнью, тем лучше она выглядит**.(с)
Следую этому высказыванию, у меня в этой сфере всё ва-ааще зашибись)))))))

poem

(7) Улицы города манят опять
Им на тебя и меня наплевать!
Надо идти только не потерять,
Будь собой, будь собой!

news

(8) Цены на смартфоны в России оказались самыми дешевыми в мире
Цены на смартфоны в России оказались самыми дешевыми в мире, об этом сообщает Hi-tech.Mail.ru. Проанализировав цены на устройства в России и в других странах, издание выяснило, что в некоторых случаях, при пересчете по актуальному курсу, разница может составить более 60%.

advertising

(9) в наличии!!!✔ Оплата на приват карту!!! ✔ Доставка 1 -2 дня!! ✔ КАК ЗАКАЗАТЬ ЖМИ СЮДА

(10) Набираю в Свою КОМАНДУ Координаторов Avon! Свободный график работы, позволяющий совмещать ДОПОЛНИТЕЛЬНЫЙ ДОХОД с основной работой, маленькими детьми или учебой!!! Бесплатный старт!!! Обучение от Лидера!!! ✨Не упускай свою возможность!!! Присоединяйся!!!!

autogen

(11) Презентация на тему: "Образ Чичикова в поэме Н.В.Гоголя «Мертвые души».". Скачать бесплатно и без регистрации.
(12) 'Викторина первоклассника (2508)' купить настольные игры | Лабиринт

link_header

(13) Они не знали, как их бульдог отреагирует на появление в доме маленького котенка. Но такого они точно не ожидали
(14) Galaxy S6 против iPhone 6 - iPhone 6 против Galaxy S6

foreign

(15) Друзі, пишіть ваші замовлення в смс або під фотографіями, аби я могла оперативно виконати їх і нікого не забути!!! Наступні замовлення виконаю на п'ятницю-понеділок!

other

(16) 2015.04.11 17.34.13 66ea80f755.720.mp4
(17) [f

2. Примеры текстов некоторых смешанных категорий

author/citation (тексты, внешне напоминающие цитаты, но не обнаруживающие дубликатов в ручном поиске**)**

(18) Что бы не случилось, у вас есть вы. И с собой можно делать много разнообразных полезных вещей, типа развития. Можно умереть в 20, а быть похороненным в 70. Ищите смыслы в себе и вы никогда не будете одиноки.

author/other (тексты, состоящие преимущественно из имен собственных, либо слишком короткие и не несущие смысла**)**

(19) Лерочка,Никуша и Анютка!

author/advertising (авторские рекламные тексты, имеющие лингвистическую ценность**)**

(20) Друзья, никому не нужен фотик Никон д3100? Идеальное состояние, юзан 2 раза, с ремнем , всеми документами , сумкой , фильтром и прочим 😊😊😊

advertising/partial_author

(21) Подруга попросила помочь продать вещи, парфюмерию ,украшения и много чего еще. За репост спасибо

"Всем привет! Так вышло, что у меня в шкафу находится целая куча классной одежды - половина абсолютно новые! Продаю всё в связи с тем, что уменьшилась

на 2 размера))) Приглашаю в гости на примерку! Уверена, найдете для себя красивые вещи по душе и по очень доступной цене. "

citation/poem (стихотворения, выписанные в строчку, но с выделением строк какими-то иными способами**)**

(22) Тем проще, тем легче ее перейти,—Там эти же рощи и озими эти ж... Ты просто ее не заметишь в пути, В беседе с ушедшим — ее не заметишь.

citation/other (цитаты с элементами авторского, но не имеющего лингвистической ценности текста**)**

(23) Поддерживаю флешмоб и приглашаю вас. Идея состоит в том, чтобы заполонить социальные сети произведениями искусства в противовес деградации общественности. Каждый, кто поставит "лайк", получит имя художника, произведение которого вы должны разместить на странице. Мне досталась вот эта: Уильям Тёрнер "Дождь, пар и скорость"