Рецензия на статью «Меры расстояния для определения авторства древнегреческих текстов»

Общий обзор работы:

Работа посвящена тестированию стилометрических метрик, основанных на векторизации текста через частотности токенов и измерении расстояний между векторами (условно такие метрики можно назвать Delta-подобными или барроузианскими), на корпусе античных текстов. Кроме собственно сравнительного анализа способности метрик классифицировать автора на античных памятниках с (относительно) известным авторством, в работе присутствует и бонусная часть, своего рода «вишенка на торте» — применение наиболее успешных метрик для получения новых аргументов в вопросах действительно спорного авторства (раздел «Псевдо-Исократ, псевдо-Плутарх, псевдо-Платон»).

Сильные стороны работы:

1. Перед нами один из самых ясно написанных текстов о современных стилометрических методах, созданных на русском языке. Как ценитель популяризации вычислительной филологии не могу не восхититься внятностью и изяществом слога без напускного наукообразия.
2. Очень ценны рассуждения о континуальности и небинарности понятия авторства — эту тонкость часто упускают из вида цифровые гуманитарные исследователи с их профдеформационной склонностью к бинаризации всего, что можно бинаризовать.
3. Удачное и редкое сочетание глубокого знакомства с гуманитарной предметной областью (видно, что автор — специалист-античник) и четкого понимания математических основ стилометрии.
4. Отрадно, что присутствует оценка статистической значимости различий между показателями качества определения авторства для различных метрик. Это всегда отрезвляет, избавляет от некоторых поспешных выводов, основанных на случайности, и делает итоговый результат сильно более убедительным.
5. Красивая кольцевая композиция: статья начинается с тезисов выдающегося немецкого античника Фридриха Бласса об основаниях и ограничениях для атрибуции авторства, затем автор переходит к современной стилометрии и ее тестированию на античных текстах, а в конце снова возвращается к наблюдениям и выводам Бласса об авторстве «Демоника», с которыми явно согласуются результаты стилометрических экспериментов.

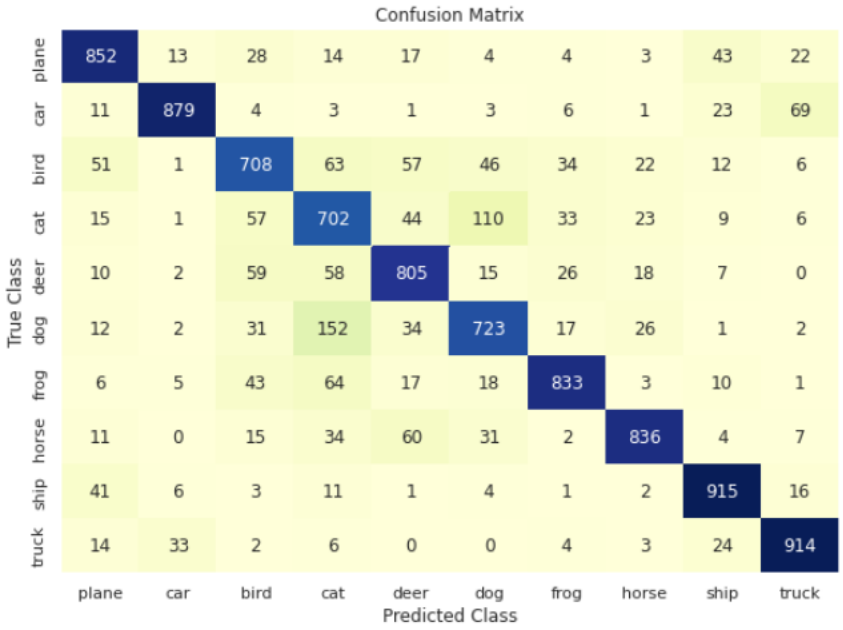
Научный контекст работы

Самим своим рождением стилометрия очевидным образом обязана развитию классической филологии и в частности платоноведения в конце XIX века ([Lutosławski, 1897], [Dittenberger, 1881]). Тем не менее сегодня существует не так много работ, где

исследовалось бы качество современных (барроузианских) стилометрических метрик на древнегреческих текстах. Гораздо чаще эти метрики тестируются и верифицируются на текстах национальных литератур нового и новейшего времени ([Eder, 2011], [Rybicki, Eder, 2011], [Jannidis, Lauer, 2014], а также многочисленные ссылки, уже приведённые в статье, вроде [Hoover, 2004]), либо на латинских памятниках (снова [Eder, 2011], а также [Eder, 2014], [Eder, 2015]). Тем ценнее работа, которая возвращает стилометрию, перевооруженную методами из XXI века, уже неоднократно подтвердившими свою универсальную (не-ситуативную, не ad-hoc) работоспособность, к ее классическому (во всех смыслах) объекту — античным греческим текстам.

Рекомендации и замечания

1. Одним из интереснейших выводов работы представляется удивительно плохой результат в определении авторства для манхэттенского расстояния со стандартизацией, то есть, собственно, для классической метрики Delta, которая по-прежнему является своего рода «plain vanilla опцией» для стилометрии и используется очень часто. Delta в этой работе проигрывает не только альтернативным метрикам вроде косинусного расстояния или Canberra distance, но и манхэттенскому расстоянию без стандартизации, то есть фактически более примитивной версии самой себя, которая должна бы работать плохо из-за избыточного влияния нескольких гиперчастотных слов (привет закону Ципфа) и плохого учета слов малочастотных. Именно чтобы снизить избыточный эффект сверхчастотных слов, Барроуз и ввел стандартизацию. А тут вдруг оказывается, что все зря. Возможно, этот факт стоит проверить и проанализировать подробнее? Не говорит ли это о том, что для античных авторов стилеметрический сигнал хорошо виден именно в сверхчастотном топе из 3–5 слов? Это было бы интересно проанализировать.
2. В Таблице 5 на стр. 10 представлены ошибки классификации для трех авторов. Фактически перед нами фрагмент матрицы ошибок (confusion matrix) классификатора. Почему бы не изобразить матрицу ошибок целиком? Это очень устоявшийся формат демонстрации того, какие классы склонны мешаться с какими при анализе работы классификатора. Традиционно матрицу ошибок рисуют в виде таблицы с раскраской тепловой карты:



Кажется, что 17 авторов в такой формат вполне уместятся.

3. На стр. 5 говорится об «использовании словоформ или трехсложных энграм». Однако из дальнейшего текста понятно, что слоговой сегментации в работе не происходило, а речь идет об обычных символьных n-граммах (character ngrams). Использование определения «трехсложный» здесь вводит читателя в заблуждение. Кажется, что «трехсимвольных энграм» будет точнее.
4. Далее те же n-граммы несколько раз называются «трехбуквенными». Однако если для их получения использовался базовый функционал stylo (выбор MFC вместо MFW при n = 3), то это тоже не совсем точно. Stylo не отличает буквы от иных символов, и для строки «Аня ест» создаст среди прочего трехсимвольную n-грамму «я е», которую едва ли можно назвать трехбуквенной. Снова кажется более точным говорить о «трехсимвольных» n-граммах.
5. Возможно, так и задумано, но в полученной на рецензию статье не было аннотации, хотя такой заголовок присутствовал. Abstract очень помогает быстро понять, о чем работа, так что хорошо бы его иметь.

Резюме

В целом статья представляет несомненный интерес и достойна публикации. Однако до публикации я бы очень советовал разобраться с тем, почему так плохо работает классическая Delta, проигрывая даже нестандартизованной версии себя. После 2002 года, когда была предложена Detla, мы видели немало работ, где предлагали усовершенствования Delta за счет более умного отбора признаков или альтернативных способов измерения расстояний между векторами. Но я не припоминаю такого, чтобы улучшения достигались путем банальной примитивизации метрики и отключения стандартизации. Такой результат как будто даже несколько дискредитирует Дж. Ф. Барроуза, так что имеет смысл разобраться, почему так происходит.

Прочие замечания касаются точности отдельных формулировок, они призваны облегчить восприятие текста и не потребуют от автора никакой дополнительной исследовательской работы или обширных правок.

Д. Скоринкин, к.ф.н., координатор сети Digital Humanities в Университете Потсдама

Библиография рецензии

1. Lutosławski W. The Origin and Growth of Plato’s Logic: With an Account of Plato’s Style and of the Chronology of His Writings. : Longmans, Green and Company, 1897. 576 c.

2. Dittenberger W. Sprachliche Kriterien für die Chronologie der Platonischen Dialoge // Hermes. 1881. Т. 16. № 3. С. 321–345.

3. Eder M. Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint // 2011. № 6. С. 99–114.

4. Eder M. Stylometry, network analysis, and Latin literature // 9th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2014, Lausanne, Switzerland, 8-12 July 2014, Conference Abstracts. : Alliance of Digital Humanities Organizations (ADHO), 2014.

5. Eder M. Taking Stylometry to the Limits: Benchmark Study on 5,281 Texts from “Patrologia Latina” // Digital Humanities 2015: Book of Abstracts, University of Western Sydney. : Alliance of Digital Humanities Organizations (ADHO), 2015.

6. Hoover D. L. Testing Burrows’s Delta // Lit. Linguist. Comput. 2004. Т. 19. № 4. С. 453–475.

7. Jannidis F., Lauer G. Burrows’s Delta and Its Use in German Literary History // Distant Readings. Topologies of German Culture in the Long Nineteenth Century Studies in German Literature Linguistics and Culture. / под ред. M. Erlin, L. Tatlock. Rochester: Camden House, 2014. С. 29–54.

8. Rybicki J., Eder M. Deeper Delta across genres and languages: do we really need the most frequent words? // Lit. Linguist. Comput. 2011. Т. 26. № 3. С. 315–321.