IT термины и профессиональный сленг в современных сообществах социальных сетей

Зорина Д.В., Тимохов В.В., Бурматова К.А., Янковская Н.С., Проноза Е.В., СПбГУ
zorro.daria@yandex.ru, viktor-timohov@mail.ru, chrisburmatova@gmail.com,
yankovskaya.natasha@gmail.com, katpronoza@gmail.com

Аннотация

В данной статье рассмотрены значимость образуемых от терминов профессиональных (сленговых) слов и выражений на примере их частот употребления в социальной среде - двух публичных страниц социальной сети ВКонтакте, а также на их употреблении в различных контекстах в рамках этой сети. Результаты работы показали, что эти частоты сильно варьируются в зависимости от термина и контекста, в котором он употребляется. В дальнейшем результаты этой работы можно использовать для пополнения профессиональных словарей и для улучшения методов информационного поиска.

1 Введение

На сегодняшний день трудно отрицать, что жизнь любого человека так или иначе затронута информационными технологиями: все мы пользуемся компьютерами, смартфонами, банковскими картами и, конечно же, Интернетом. Объем информации, обрабатываемой и хранимой этой системой, колоссален, она охватывает бесконечное пространство всевозможных данных, ежедневно пополняемое все новыми и новыми единицами. Также все новыми и новыми лингвистическими единицами посредством сферы IT пополняется пространство нашего языка, стремительно развивающиеся технологии обогащают его огромным количеством новых терминов [Волоснова, 2010]: названия операционных систем, программ, языков программирования и т.д. -- неологизмы, уже плотно вошедшие в словарный запас среднестатистического пользователя именно как термины. Тем не менее, данные термины, постоянно применяемые в процессе обмена информацией, не остаются неизменными [Костина, 2014] [Посевкин., Мальнев, 2010]. Современный русский язык пестрит огромным количеством их вариаций, форм, подчас совершенно неожиданных [Галичкина, 2007]. И разнообразные формы активно используются как носителями официального терминологического языка (например, учебники, справочники, официальные документы), профессионалами, носителями профессионального неформального языка (коммуникация профессиональных сообществ как на уровне социальных сетей и профессиональной Интернет-коммуникации в целом, так и не совсем официальных учебников, справочников, вопросников и т.д.), а также обычными пользователями в различном контексте.

Профессиональная терминология особенно важна для изучения. Она облегчает коммуникацию в среде специалистов, позволяя им проще понимать друг друга и подстраивать термины под особенности своего языка. Такого рода терминология не только объединяет профессиональное сообщество в (и особенности их коммуникации), но и позволяет новичкам приобщиться к ней. Более того, существенное количество востребованной информации приходится именно на тексты с такой терминологией (скажем, информацию о текущих изменениях библиотек, версий языков и т.д. скорее можно найти в такого рода текстах) в отличие от базовой информации учебников и справочников.

Предположением работы является тот факт, что, изучив соотношение официальных названий и их сленговых аналогов, а также их контекста - слов, используемых в тех же блоках текста и позволяющих найти информацию, связанную с данным термином, можно будет изучить способы улучшения эффективности поисковых запросов путём определения, какую именно фразу или слово следует использовать для получения определенной области знаний о предмете.

В фокус внимания нашего исследования попадает вопрос о соотношении жанра публикации и наиболее частотными вариантами того или иного термина (официального или сленгового).

2 Методы

Для исследования различия в употреблении официальных и сленговых терминов, нам необходимо было собрать информационную базу, на основе которой можно было бы сделать выводы о частоте употребления различных форм выражений и их контекста. В данной работе были использованы две публичные страницы социальной сети ВКонтакте, из которых были извлечены тексты всех публикаций (постов) до 08.10.2017 и до 100 комментариев пользователей на один пост.

Для очистки полученных данных были реализованы следующие функции: удаление ссылок, пунктуации, информации о пользователях и приведение к нижнему регистру. Произведенная очистка была минимальная, так как среди коротких и редко встречающихся слов также были обнаружены полезные для исследования данные.

После обработки извлеченной информации, при помощи простого алгоритма был сформирован частотный словарь, с которым работала наша исследовательская группа (В дальнейшем возможно дополнение словаря информацией из других источников, также возможна и его очистка от лишних данных, которые могут накапливаться в процессе роста словаря).

Для того чтобы получить наиболее часто встречающиеся и близкие по смыслу слова в нашей предметной области, была обучена дистрибутивно-семантическая модель Skip-gram [Mikolov, Chen, Corrado, Dean, 2013] на основе технологии Word2vec [Goldberg, Levy, 2014], которая, в свою очередь, была обучена на исходных данных частотного словаря.

2.1 Выбор терминов для изучения

Использовались два параллельных способа установления терминов для более подробного изучения: пользовательский опыт и наибольшая частота их употребления, которая говорит о распространенности, а значит, и о значимости этих терминов для текущего исследования.

2.2 Использованные инструменты

- язык программирования Python;
- модуль для работы с векторными моделями gensim.

3 Результаты

Из полученного частотного словаря были сформированы примерный список синонимов (полный (Приложение 1) и краткий варианты (Таблица 1)) и список, содержащий процентное соотношение упоминаний официальных терминов и их аналогов из компьютерного сленга (Таблица 2), а также составлена таблица, отражающая частоту использования их в контексте различного характера (Таблица 3).

Также стоит уточнить, как именно мы определяем компьютерный сленг в данном исследовании. Он включает в себя множество терминов, используемых как профессионалами в IT-сфере, так и посредственными пользователями. Существует множество способов формирования новых терминов из заимствований, а также пополнения профессионального сленга, но считали основными 4 из них: калька (*Java - Джава*), полукалька (*Windows - Винда*), перевод (*Ruby - Рубин*) и фонетическая мимикрия (*Android - Андрюха*), дополнительные способы формирования терминов (прежде всего, неофициальных) лежат в области языковой игры, позволяющей адаптировать иноязычную терминологию (в данном случае – англоязычную) к особенностям функционирования русского языка.

Официальными (не сленговыми) терминами мы считали те, которые либо содержатся в словарях, либо являются официальными названиями (например, имеют определенное название, зафиксированное в ГОСТах и официальной технической документации).

Табл. 1. Список сленговых слов

Общепринятое название	**Сленговые вариации**
Windows	Винда, виндовс, win, вин, шиндовс, венда, шиндоуз
Java	Джава, ява, жаба
Ubuntu	Убунта, бубунта, хубунту, убунт
Linux	Линь, линух, линус, линя
Python	Питон, пайтон, петон, путон, питоний
Android	Андроид, ведро, андройд, ведроид, дроид, андрюха, андрюша
C++	Плюсы, cpp, сисиплюс, плюсики
Kaspersky	Касперский, каспер, касперыч

Табл. 2. Процентное соотношение частоты использования официальных терминов и их аналогов из компьютерного сленга (%)

	Оригинал	**Синонимы**
Windows	29,6	70,4
Java	76,3	23,7
Ubuntu	57,5	42,5
Linux	71,9	28,1
Python	36,5	63,5
Android	28,5	71,5
C++	72,9	27,1
Kaspersky	5,7	94,3

Эти таблицы (табл. 1 и 2) показывают, что частоты использования официальных и сленговых форм для приведенных названий разнятся, причем для одних терминов преобладающей является частота использования официального названия (Java, Ubuntu, Linux, C++), а для других - сленгового (Windows, Python, Android, Kaspersky). При этом для некоторых терминов та или иная частота составляет менее 10 от общего числа упоминаний (Kaspersky).

Табл. 3. Процентное соотношение упоминаний оригинальных (ориг) названий и синонимов (син) в контексте различного характера (%)

		Советы, вопросы	Статьи, новости	Юмор
Windows	Ориг	94	2	4
	Син	97	1	2
Linux	Ориг	81	8	11
	Син	82	1	17
Ubuntu	Ориг	87	4	9
	Син	83	1	16
Python	Ориг	78	2	20
	Син	68	-	32
JavaScript	Ориг	95	4	1
	Син	92	1	7
Kaspersky	Ориг	50	42	8
	Син	66	13	21
C++	Ориг	50	14	36
	Син	78	4	18

В таблице 3 приведено сравнение частот использования некоторых названий (их официальных и сленговых форм) в 3 видах контекста: технические советы (вопросы), статьи (новости) и юмор. При этом все термины (все их формы) наиболее часто употреблялись именно в контексте технического совета (вопроса), и частоты использования официального названия и его аналога из компьютерного сленга в данном контексте были примерно равны. Скорее всего, такие результаты связаны с особенностями публичных страниц ВКонтакте, материал которых был изучен, и их контингента. По большей части пользователи общаются в них для решения каких-либо вопросов и задач, возникших в их работе/проекте. При этом сами администраторы этих страниц производят контент, связанный с профессиональным юмором и с новостями из IT-сферы.

В таблице 4 представлены примеры употребления выбранных названий и их синонимов для выявленного типа контекста.

Табл. 4. Примеры контекста

	Примеры для оригинального термина	**Примеры для сленга**
Советы, вопросы	1.Интересно, windows хоть раз находил способ разрешения проблемы как он заявляет? 2.Ставь lubuntu. Ubuntu тяжелее винды, а lubuntu лёгкий дистрибутив В Windows я играюсь, что весьма редко, Photoshop с AI, раньше пользовался iTunes, ведь под Linux нативного нету, а linux для работы, веб-разработки очень даже мне подходит.	1.Учи жабу 2.На счёт того, что меня "восхищает" в ведроиде — это отсутствие любых препонов. 3.Я просто не понимаю зачем нужна сейчас ява.
Статьи, новости	1.Многие пользователи Windows считают операцию одновременного нажатия трех клавиш неудобной, при этом большинство привыкло к ней за годы существования различных версий Windows. 2.Инструкция по поиску и устранению ошибок в коде для языка Python 3.Неизвестные хакеры выложили в общий доступ инструменты, которыми пользовались спецслужбы для взлома iPhone, а также сотен смартфонов на Android.	1.В русскоязычной среде принято говорить "Питон". 2.Ищем программиста на питоне 3.Особенно если там установлен какой-нибудь эмулятор виндоус - потенциальная дыра в безопасности.
Юмор	1.Чойто java - зло? Я java разработчик ващет, салаги.. 2.Linux — только одна фраза: «Если на свой компьютер вы сумели поставить операционную систему семейства Linux, то мы не сомневаемся, что и с настройкой Wi-Fi проблем у вас не возникнет!» 3.Это как оккультизм в java	1.На словах ты лев толстой, а в плюсах ты джун простой. 2.У меня были однокурсники, которые путали паскаль с плюсами на 3 курсе 3. Алло, я сейчас не могу говорить, я пользуюсь Линуксом

4 Вывод и перспективы

Учитывая методику поисковых машин, использующих по большей части ключевые слова, привлекающие разные результаты поиска, и результаты, полученные в процессе исследования, можно сделать вывод, что более глубокое исследование сленговых синонимов терминов, в том числе способов их образования, а, главное, функционального контекста -- информационной оболочки, их содержащей -- открывает перспективы установления специфических пользовательских вариантов и синонимов для более широкого круга терминов, что поможет улучшить результативность поисковых систем. Это улучшение может происходить двумя путями: 1) изменением алгоритмов, чтобы поиск происходил не только по традиционным ключевым словам, но и по их формальныи и неформальным синонимам; и 2) добавлением новых баз данных, содержащих сленговые выражения, в поисковые системы или добавлением функционала помощи пользователю, совершающему поиск.

Список литературы

Волоснова, Ю.А. 2010. *Образование неологизмов (на примере компьютерного сленга).* Казанский социально-гуманитарный вестник.

Галичкина, Е.Н. 2007. *Людический Потенциал компьютерного сленга как лингвокультурного феномена.* Вестник государственного областного университета. Серия: лингвистика - №2. - С.108-114.

Goldberg Y., Levy O. 2014. *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method.*

Костина, Е.В. 2014. *Феномен трансформации компьютерной лексики в русском языке: от термина к сленгу*. Вестник гуманитарного факультета Ивановского государственного химико-технологического университета - №6. - С.39-45.

Mikolov, T., Chen, K., Corrado, G., Dean, J. 2013. *Efficient estimation of word representations in vector space.* In Proceedings of Workshop at ICLR.

Посевкин Р.В., Мальнев М.А. 2010. *Компьютерные жаргонизмы: По материалам IX научно-практической конференции ППС ВПИ.* г. Волжский, С.190-192.

5 Приложение 1 Примеры контекста. Полный список сленговых слов, используемых в исследовании

Табл. 5. Примеры контекста Полный список сленговых слов, используемых в исследовании

Общепринятое название	**Сленговые вариации**
Windows	Винда, виндовс, win, вин, шиндовс, венда, шиндоуз
JS, JavaScript	Джаваскрипт, яваскрипт, жабаскрипт
Java	Джава, ява, жаба
Ubuntu	Убунта, бубунта, хубунту, убунт
Linux	Линь, линух, линус, линя
Basic	Басик, васик
iPhone	Ипхон, ифон, ойфон, гейфон, яблофон
C#	Шарп, сишарп
Python	Питон, пайтон, петон, путон, питоний
PHP	Пхп, похапе, похапэ
Android	Андроид, ведро, андройд, ведроид, дроид, андрюха, андроед
Ruby	Руби, рубин
HTML	Хтмл, хтмль
OS	ОС, системы, ось,, операционки
Вирус	Вирусняк, малварь
C++	Плюсы, сипипи, сиплюс, плюсики, сиплюсы
Macintosh	Мак, макинтош
Macos	Макось, макос
CSS	ЦСС, сиэсэс
Компьютер	Комп, компуктер , кампуктер, кампухтер, кудахтер, камп
iOs	Айос, иос, айось
Kaspersky	Касперский, каспер, касперыч