

Automatic classification of BootCat'ed corpora

Serge Sharoff

Partly funded by Project TTC
Terminology Extraction, Translation Tools and Comparable Corpora
(FP-7, STREP-248.005)

24 June 2013

Outline

- 1 Analysing BootCat'ed corpora
 - Keywords for collection
 - Clustering and topic modeling
- 2 Measuring text similarity
 - Traditional genre classification
 - Text property annotation

Collecting texts on renewable energy

- Keywords from Wikipedia's Renewable Energy category

fossil fuel	化石燃料	ископаемое топливо
power station	发电厂	электростанция
hydroelectricity	水力发电	гидроэнергетика
photovoltaics	太阳能光伏	фотоэлектричество

- Corpora from BootCat (Bing)

	En	Ru	Zh(CN)	Zh(TR)
URLs:	5762	5991	674	870
Words (MW):	6.5	5.8	1.9	1.68

Assessing the composition by keywords

7467 renewable energy
4352 wind turbine
3973 fossil fuel
3127 greenhouse gas
3049 natural gas
2539 wind farm
2320 solar energy
2265 energy efficiency
1994 carbon dioxide
1920 solar cell
1782 wind energy
1722 generate electricity
1559 solar patch
1533 electricity generation

5629 источник энергия 'energy source'
4550 окружающий среда 'environment'
2754 электрический энергия 'electricity'
2710 солнечный батарея 'solar cell'
2274 солнечный энергия 'solar energy'
2106 природный газ 'natural gas'
1994 тепловой энергия 'thermal energy'
1870 возобновлять источник 'renewable energy'
1561 производство электроэнергии 'electricity generation'
1508 возобновлять источник энергия 'renewable energy source'
1439 изменение климат 'climate change'
1401 парниковый газ 'greenhouse gas'
1315 альтернативный источник 'alternative source'
1289 энергия ветер 'wind energy'



LEEDS

Classification of BootCat

Self-Charging Solar Cells: Better Than Batteries? - Forbes - Google Chrome

UKWAC: conformant% w... solar cell Self-Charging Solar Cells: ...

www.forbes.com/sites/williampentland/2013/06/12/self-charging-solar-cells-better-than-batteries/

Forbes

New Posts

Popular
Fast Tech 25

Lists
Highest-Paid Athlete: 'Buycott' App In Action

Video
'Buycott' App In Action


Search

ENERGY | 6/12/2013 @ 3:46PM | 20,810 views

Self-Charging Solar Cells: Better Than Batteries?

+ Comment Now + Follow Comments

A research team at the [University of Wisconsin](#) in Madison has demonstrated the viability of a design for solar panels that can simultaneously generate and store electrons harvested from sunlight in a single device.



First Solar

Like other solar panels, the new solar cells would use some electrons from light created as electricity. Unlike other solar cells, the panels would also store electrons on zinc oxide nanowires coated with polyvinylidene fluoride polymer (PVDF). The energy stored during the day could be used to run the lights at night or on cloudy days.

The PVDF has a high dielectric constant, which releases the stored energy through the nanowires when the solar cell stops harvesting

210

f Share

227

Twitter

11

in Share

0

Submit

3

+1

0

Most Read on Forbes

NEWS People Places Companies

Obamacare Is Turning Walmart Workers Into Temps +181,996 views

The New iOS 7 Design: What Works, What Needs Work +51,723 views

Why Praise For The PS4 And Criticism Of Xbox One Are Vastly Overdone +49,002 views

Why Xbox One And PS4 Have Me Worried For The Future Of Games +38,550 views

Democratic Congressman: 'Not Fair' To Subject Congress To Obamacare Just Like Everyone Else +37,953 views

COOKIES ON FORBES

UNIVERSITY OF LEEDS

19.9%-efficient ZnO/CdS/CuInGaSe₂ solar cell with 81.2% fill factor - Repins - 2008 - Progress in Photovoltaics: Research and Appl

UKWAC: conformat% w... 19.9%-efficient ZnO/CdS/ x

onlineibrary.wiley.com/doi/10.1002/pip.822/pdf

Progress in Photovoltaics: Research and Applications
Volume 16, Issue 3, Article first published online: 14 FEB 2008
Abstract | References | Cited By

UNIVERSITY OF LEEDS WILEY ONLINE LIBRARY

PROGRESS IN PHOTOVOLTAICS: RESEARCH AND APPLICATIONS
Prog. Photovolt: Res. Appl. 2008; **16**:235–239
Published online 14 February 2008 in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/pip.822

Research SHORT COMMUNICATION: ACCELERATED PUBLICATION

19.9%-efficient ZnO/CdS/ CuInGaSe₂ Solar Cell with 81.2% Fill Factor[†]

Ingrid Repins^{1,*†}, Miguel A. Contreras¹, Brian Egas¹, Clay DeHart¹, John Scharf¹, Craig L. Perkins², Bobby To¹ and Rommel Noufi¹

¹National Renewable Energy Lab, MS 2219, CO, USA
²National Renewable Energy Lab, MS 3218, CO, USA

We report a new record total-area efficiency of 19.9% for CuInGaSe₂-based thin-film solar cells. Improved performance is due to higher fill factor. The device was made by three-stage co-evaporation with a modified surface termination. Growth conditions, device analysis, and basic film characterization are presented. Published in 2008 by John Wiley & Sons, Ltd.

KEY WORDS: CIGS; thin film solar cells; record efficiency; fill factor; recombination; diode quality; saturation current; surface

Received 20 November 2007; Revised 9 January 2008

INTRODUCTION

Record-efficiency devices are of interest for several reasons. First, they provide a proof of concept for developing products that require higher power per area, lower cost per watt, or higher watts per kg. Perhaps more importantly, understanding the sensitivities and physical mechanisms that lead to improved

stage co-evaporated CIGS, chemical-bath-deposited (CBD) CdS, sputtered resistive/conductive ZnO bi-layer, e-beam-evaporated Ni/Al grids, MgF₂ anti-reflective coating, and photolithographic device isolation. These device layers have been described in previous publications.^{1–5} Figure 1 shows logged CIGS deposition data from the 19.9% device, M2992. The graph reflects slight optimizations to deposition

Testing with topic models: English

Table, Equipment, Market, Consumption, Capacity, Production, Industry, Generation, Distrib
earth, atmosphere, dioxide, surface, cool, cause, warming, fluid, radiation, methane, reservoir
reactor, Nuclear, uranium, radioactive, barrel, mine, Uranium, fission, cent, Petroleum, reserv
ocean, wave, OTEC, Ocean, tide, Intel, Tidal, Wave, marine, Hawaii, offshore, Conversion, d
Commission, Public, shall, bill, Utility, credit, Federal, contract, FERC, eligible, District, regu
distribute, consumer, distribution, network, peak, meter, period, investment, value, datum, a
speed, rotor, blade, field, magnetic, shaft, circuit, wire, engine, transformer, phase, connect,
cogeneration, hydrogen, ethanol, engine, wood, combustion, Biomass, boiler, burn, crop, con
Green, News, Hydro, Business, India, Stock, Development, Sustainable, Geothermal, Alternat
river, hydroelectric, hydro, reservoir, fish, River, head, blade, hydroelectricity, Hydro, Hydrop
post, read, bill, want, look, article, problem, money, green, really, question, link, warming, ne
announce, billion, Tags, investment, megawatt, Kansas, expect, sign, April, News, green, Cal
sustainable, policy, economic, sector, reduction, Development, management, national, interna
module, light, silicon, inverter, watt, sunlight, hour, roof, saving, device, appliance, save, film
Program, National, Department, California, Center, Association, Resources, Public, Informati

Testing with topic models: Russian

финансовый (financial), государство (state), инвестиция (investment), мера (measures), до (proportion), политика (politics/strategy), правительство (government),

вот (this/yeah), кто (who), да (yes), сделать (do), сейчас (now), говорить (say), там (there), ни (neither), ли (would), надо (should), просто (simply), сам (-self), знать (know)

конференция (conference), выставка (exhibition), устойчивый (sustainable), наука (science), энергосбережение (energy saving), энергоэффективность (energy efficiency)

геотермальный (geothermal), биомасса (biofuel), водород (hydrogen), отходы (waste), возобновляемый (renewable), ветроэнергетика (wind generation), топливный, вэу (wind farm), море (sea), глобальный (global), океан (ocean), растение (plant), лес (forest), загрязнение (pollution), потепление (warming), природа (nature), парниковый (greenhouse)

плотина (dam), добыча (production), сооружение (installation), водохранилище (reservoir), месторождение (deposit), запас (resource), очистка (purification), сточный (sewage)

аккумулятор (battery), воздушный (air), насос (pump), емкость (capacity), нагрузка (load), рисунок (drawing), ротор (rotor), конструкция (design), вал (shaft), корпус (body)

тэц (CHP), энергосистема (grid), газотурбинный (gas turbine), когенерация (cogeneration), киловольт (kV), нагрузка (load), котельная (boiler station)

паровой (steam), котел (boiler), силовой (power), трансформатор (transformer), ГОСТ (GOST), частота (frequency), пар (steam), линия (line), преобразователь (converter), передача (transmission), реактивный (reactive), переменный (alternating), провод (wires)

реферат (essay), реактор (reactor), движение (movement), ядро (nucleus), наука (science), поле (field), атом (atom), реакция (reaction), нейтрон (neutron), физика (physics), планета (planet), магнитный (magnetic)

новость (news), ноутбук (laptop), процессор (processor), комментарий (comment), компьютер (computer)

Outline

- 1 Analysing BootCat'ed corpora
 - Keywords for collection
 - Clustering and topic modeling
- 2 Measuring text similarity
 - Traditional genre classification
 - Text property annotation

Brown, LOB, LCMC 500 samples, 2000 words in each, belonging to 15 genres: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure ... R) Humour

Defining genre categories

Brown, LOB, LCMC 500 samples, 2000 words in each, belonging to 15 genres: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure ... R) Humour

BNC about 4,000 texts with classification into 70 genres (ac.med, ac.tech, non-ac.tech, news...), medium (book, periodical, ephemeral, ...), audience, ...

Defining genre categories

Brown, LOB, LCMC 500 samples, 2000 words in each, belonging to 15 genres: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure ... R) Humour

BNC about 4,000 texts with classification into 70 genres (ac.med, ac.tech, non-ac.tech, news...), medium (book, periodical, ephemeral, ...), audience, ...

BL for fiction Adventure stories, Detective stories, Picaresque literature, Robinsonades, Sea stories, Spy stories, Thrillers, Allegories, Didactic fiction, Fables, Parables, Alternative histories, Dystopias, Bildungsromane, Arthurian romances, ...

Defining genre categories

Brown, LOB, LCMC 500 samples, 2000 words in each, belonging to 15 genres: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure ... R) Humour

BNC about 4,000 texts with classification into 70 genres (ac.med, ac.tech, non-ac.tech, news...), medium (book, periodical, ephemeral, ...), audience, ...

BL for fiction Adventure stories, Detective stories, Picaresque literature, Robinsonades, Sea stories, Spy stories, Thrillers, Allegories, Didactic fiction, Fables, Parables, Alternative histories, Dystopias, Bildungsromane, Arthurian romances, ...

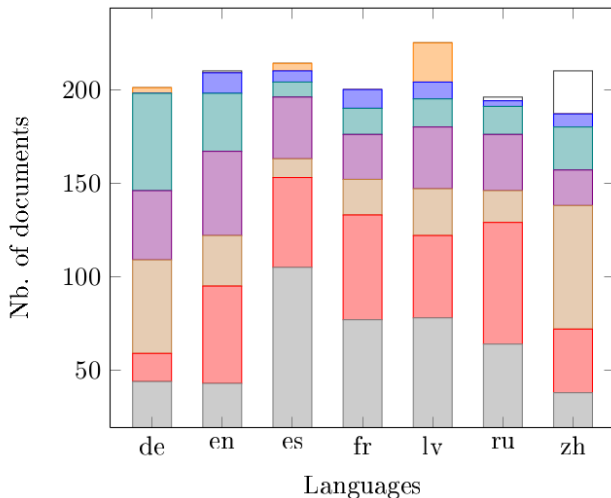
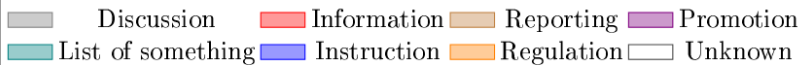
Adamzik (1995) 4,000 Textsorten

List from Adamzik (1995)

<i>Abänderungsantrag</i>	<i>Abrüstungsverhandlungen</i>	<i>Adversaria</i> [vor Augen liegende
<i>Abbestellung</i>	<i>Absage</i>	Kladde mit ungeordneten Kon-
<i>Abbruchgenehmigung</i>	<i>Absatz</i>	zepten, Notizen]
<i>Abdankungserklärung</i>	<i>Absatzgarantie</i>	<i>Agenda</i> [Notizbuch]
<i>Abecedarium</i>	<i>Abschiedsbrief</i>	<i>Agende</i> [Kirch]
<i>Abendblatt</i>	<i>Abschiedsgespräch</i>	<i>Agentenroman</i>
<i>Abendgebet</i>	<i>Abschiedsrede</i>	<i>Agenturbericht</i>
<i>Abendgespräch</i>	<i>Abschilderung</i>	<i>Agenturmeldung</i>
<i>Abendnachrichten</i>	<i>Abschlußarbeit</i>	<i>Agitpropstück</i>
<i>Abendprogramm</i>	<i>Abschlußbesprechung</i>	<i>Ahnenprobe</i>
<i>Abendzeitung</i>	<i>Abschlußbilanz</i>	<i>Ahnentafel</i>
<i>Abenteuerroman</i>	<i>Abschlußgespräch</i>	<i>Akkordzettel</i>
<i>Aberkennung</i>	<i>Abschlußrechnung</i>	<i>Akkreditiv</i> [Beglaubigungsschrei-
<i>Abfahrtsplan</i>	<i>Abschlußzeugnis</i>	ben eines Diplomaten]
<i>Abfindungserklärung</i>	<i>Abschnitt</i>	<i>Akquisitionsliste</i>
<i>Abgabebewilligung</i>	<i>Abschrift</i>	[Anschaffungsliste]
<i>Abgabeordnung</i>	<i>Abschußliste</i>	<i>Akte</i>
<i>Abgangsmeldung</i>	<i>Abschußplan</i>	<i>Aktenband</i>
<i>Abgangszeugnis</i>	<i>Abschwörungsformel</i>	<i>Aktenfaszikel</i>
<i>Abgeordnetenrede</i>	<i>Absichtserklärung</i>	<i>Aktenheft</i>
<i>Abgesang</i> [im Meistersang]	<i>Absolutorium</i> [Reifezeugnis;	<i>Aktennotiz</i>
<i>Abhandlung</i>	österr.: Bestätigung einer Hoch-	<i>Aktenstück</i>
<i>Abhang</i> [ind. Hymne]	schule über erbrachte Leistungen]	<i>Aktenvermerk</i>
<i>Abhörverbot</i>	<i>Abstammungsklage</i>	<i>Aktie</i>
<i>Abiturientenzeugnis</i>	<i>Abstammungsnachweis</i>	<i>Aktiengesetz</i>
<i>Abiturzeugnis</i>	<i>Abstammungsurkunde</i>	<i>Akzept</i> [Bank]
<i>Abkommen</i>	<i>Abstimmungsunterlagen</i>	<i>Akzessionsliste</i> [Verzeichnis von

Functional Genre Classes (FGC)

- ① **information** (catalogues, glossaries, home pages)
- ② **instruction** (how-tos, FAQs, tutorials)
- ③ **promotion** (adverts, shops, political pamphlets?)
- ④ **recreation** (fiction and popular lore)
- ⑤ **regulations** (laws, small print, contracts)
- ⑥ **reporting** (newswires, police reports)
- ⑦ **discussion**:
 - academic (research papers and monographs)
 - public (journalism and political debates)
 - everyday communication (forums, emails, diary blogs)
- ⑧ **non-text** (pages with little running text):
 - applications (Flash, Java, applets); online interfaces (query/login/purchase forms, download); linkerie (portals, link lists)



Low interannotator agreement

- double annotation on I-EN (100 pages): $\alpha = 0.70$

Low interannotator agreement

- double annotation on I-EN (100 pages): $\alpha = 0.70$
- Krippendorff (2004:241): $\alpha \geq .800$ is reliable;
 $\alpha \geq .667$ is the lowest conceivable limit

Low interannotator agreement

- double annotation on I-EN (100 pages): $\alpha = 0.70$
- Krippendorff (2004:241): $\alpha \geq .800$ is reliable;
 $\alpha \geq .667$ is the lowest conceivable limit
- only *regulation* is reliable (.93);
reporting (.57), *discussion* (.64) and *promotion* (.55) are the worst offenders

Low interannotator agreement

- double annotation on I-EN (100 pages): $\alpha = 0.70$
- Krippendorff (2004:241): $\alpha \geq .800$ is reliable;
 $\alpha \geq .667$ is the lowest conceivable limit
- only *regulation* is reliable (.93);
reporting (.57), *discussion* (.64) and *promotion* (.55) are the worst offenders
- TTC: 120 pages, 3 rounds, 8 independent annotators for English
 $\alpha = 0.50$

Text type annotation for 17 categories

- A1. To what extent does the text seek to persuade the reader to support (or renounce) an opinion or point of view?

Text type annotation for 17 categories

- A1. To what extent does the text seek to persuade the reader to support (or renounce) an opinion or point of view?
- A2. To what extent is corporate authorial responsibility indicated?

Text type annotation for 17 categories

- A1. To what extent does the text seek to persuade the reader to support (or renounce) an opinion or point of view?
- A2. To what extent is corporate authorial responsibility indicated?
- A3. To what extent is the text concerned with expressing feelings or emotions?

Text type annotation for 17 categories

- A1. To what extent does the text seek to persuade the reader to support (or renounce) an opinion or point of view?
- A2. To what extent is corporate authorial responsibility indicated?
- A3. To what extent is the text concerned with expressing feelings or emotions?
- A4. To what extent is the text's content fictional?

Text type annotation for 17 categories

- A1. To what extent does the text seek to persuade the reader to support (or renounce) an opinion or point of view?
- A2. To what extent is corporate authorial responsibility indicated?
- A3. To what extent is the text concerned with expressing feelings or emotions?
- A4. To what extent is the text's content fictional?
- A5. To what extent is the text light-hearted, i.e. aimed mainly at amusing or entertaining the reader?

Text type annotation for 17 categories

- A1. To what extent does the text seek to persuade the reader to support (or renounce) an opinion or point of view?
- A2. To what extent is corporate authorial responsibility indicated?
- A3. To what extent is the text concerned with expressing feelings or emotions?
- A4. To what extent is the text's content fictional?
- A5. To what extent is the text light-hearted, i.e. aimed mainly at amusing or entertaining the reader?
- A6. To what extent is the text's content written in an informal style?

Text type annotation for 17 categories

- A1. To what extent does the text seek to persuade the reader to support (or renounce) an opinion or point of view?
- A2. To what extent is corporate authorial responsibility indicated?
- A3. To what extent is the text concerned with expressing feelings or emotions?
- A4. To what extent is the text's content fictional?
- A5. To what extent is the text light-hearted, i.e. aimed mainly at amusing or entertaining the reader?
- A6. To what extent is the text's content written in an informal style?

Rating Levels:

- | | |
|-----|---------------------------|
| 0 | none or hardly at all; |
| 0.5 | slightly; |
| 1 | somewhat or partly; |
| 2 | strongly or very much so. |

(Forsyth, Sharoff, 2013)

Categories continued

- A7. To what extent does the aim of the text seem to be to teach the reader how to do something (e.g. a tutorial)?
- A8. To what extent does the text appear to be a newswire story?
- A9. To what extent does the text lay down a contract or specify a set of regulations?
- A10. To what extent does the text represent spoken discourse?
- A11. To what extent is the text is written from a first-person point of view?
- A12. To what extent does the text promote a commercial product or service?
- A13. To what extent is the text intended to promote a political movement, party, religious faith or other cause?
- A14. To what extent does the text require background knowledge of a specialized subject area in order to be comprehensible?

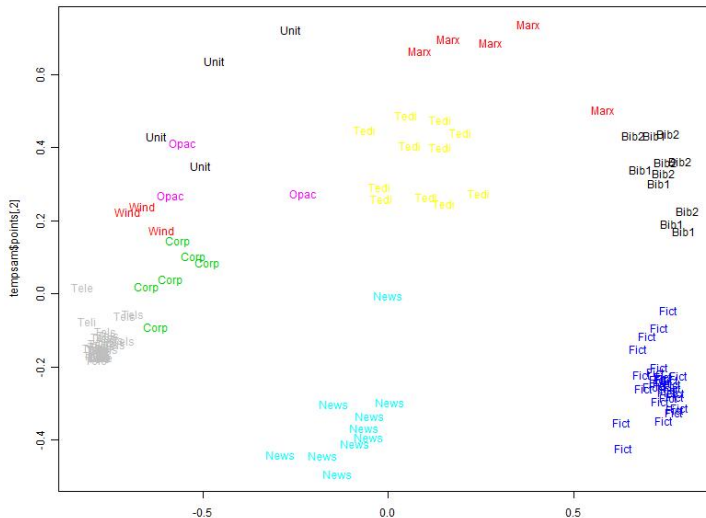
Annotation examples

A	B	C	D	E	F	G	H	I	J	K	L	M
Text	polemic	polemic	polemic	MEDIAN	emotive	emotive	emotive	MEDIAN	fictive	fictive	fictive	MED
Bib1Amos_Prophecy_89_EN.txt	2	0	2	2	1	2	1	1	2	1	1	1
CorpHandM_Quality_EN.txt	0	1	0.5	0.5	0	0	0	0	0	0	0	0
FictFlaubertG_Salamambo_11_EN.txt	0	0	0	0	1	2	0.5	1	2	2	2	2
FictGrimmJ_Bremusicians_EN.txt	0	0	0	0	1	2	0	1	2	2	2	2
MarxMarxK_ComMan_01_EN.txt	2	2	2	2	1	1	0.5	1	0.5	0.5	0	0.5
OpacTeam_Berlin_EN.txt	1	0.5	1	1	0.5	0	0.5	0.5	0	0	0	0
TeleHTC_Manual_7_EN.txt	0	0	0	0	0	0	0	0	0	0	0	0
TeleHTC_Manual_13_EN.txt	0	0	0	0	0	0	0	0	0	0	0	0
TelsGoog_Answer_770f_EN.txt	0.5	0	0	0	0	0	0	0	0	0	0	0
TelsGoog_Answer_3024_EN.txt	0	0	0.5	0	0	0	0	0	0	0	0	0
UnitUnat_HumanRights_EN.txt	2	1	2	2	1	1	1	1	0	0	0	0
WindRiadhEt_Contrarotating_EN.txt	1	1	1	1	0	0	0	0	0	0	0	0
OpacTeam_Budapest_EN.txt	2	0.5	2	2	1	0.5	0.5	0.5	0	0	0	0
MarxMarxK_ComMan_24_EN.txt	2	2	2	2	1	1	0.5	1	0	0	0.5	0
FictFlaubertG_Salamambo_2_EN.txt	0	0	0	0	1	2	0.5	1	2	1	2	2
TediJakubowskiM_OpenTech_EN.txt	2	1	1	1	1	1	0	1	0	0	0	0
CorpApple_Environment_EN.txt	0.5	1	0.5	0.5	0	0	0	0	0	0	0	0
FictPoeE_Purloined_EN.txt	0	0	0	0	0.5	2	0	0.5	2	2	2	2

$\alpha \approx 0.60..0.90$

Multidimensional scaling: $17 \rightarrow 2$

Sammon scaling on combined 5gcorpus ratings (stress=0.0264).



temp[sam]points[1]

Different corpora, different languages, different genres

Brown, LOB, LCMC 15 genres: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure ... R) Humour

Different corpora, different languages, different genres

Brown, LOB, LCMC 15 genres: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure . . . R) Humour

Russian genres Administrative, Journalistic, Literary, Scientific

Different corpora, different languages, different genres

Brown, LOB, LCMC 15 genres: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure . . . R) Humour

Russian genres Administrative, Journalistic, Literary, Scientific

Features mixed representation of unigrams
(top 1000 words and POS tags for others)
a, about, after, all, know, good,
must. . . CC, CD, JJ, VVZ

Different corpora, different languages, different genres

Brown, LOB, LCMC 15 genres: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres, H) Miscellaneous, J) Learned, K) Fiction: general, L) Fiction: mystery and crime, M) Adventure . . . R) Humour

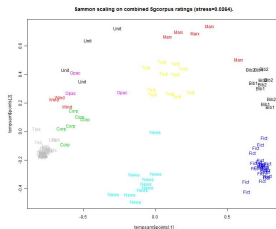
Russian genres Administrative, Journalistic, Literary, Scientific

Features mixed representation of unigrams
(top 1000 words and POS tags for others)
a, about, after, all, know, good,
must. . . CC, CD, JJ, VVZ

Machine Learning SVM regression (RBF kernel)

	5g-en	5g-de	5g-fr	5g-ru	5g-zh
Correlation X	0.92	0.93	0.92	0.95	0.95
Correlation Y	0.80	0.69	0.72	0.75	0.90

Significant features for X and Y



Negative X (,), T0, VV, perplexity, area, avoid, click, computer, contain, group, include, information, link, note, type, want, work, your

Positive X CD, DT, VVD, WP, although, come, girl, have, heart, himself, house, life, speak, tell, their, them, they, town, woman

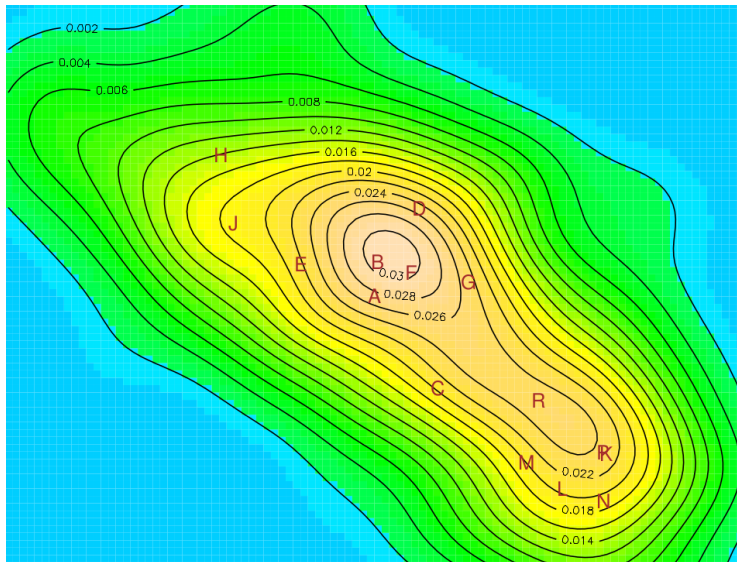
Negative Y ', JJ, NP, UH, VVD, VVG, ', almost, black, both, close, group, himself, little, mark, quite, relate, road, seem, young

Positive Y CD, WDT, actually, although, appropriate, build, come, condition, family, into, nation, principle, raise, school, social, society, tell, thing, whole, world

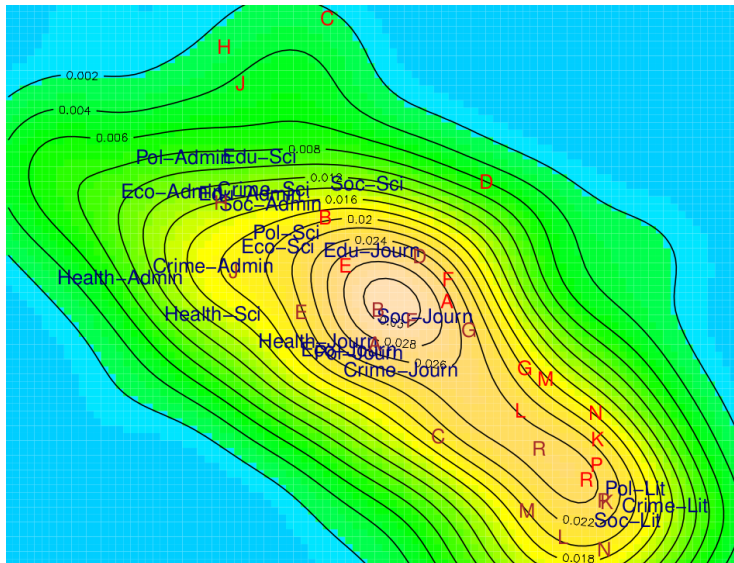
X Operative → Descriptive

Y Narrative → Argumentative

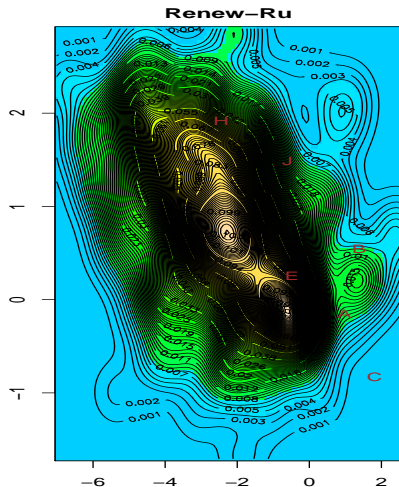
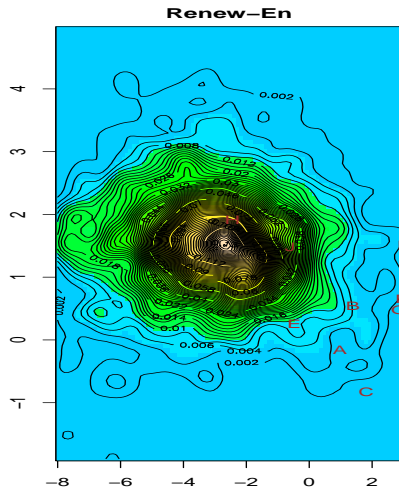
Brown texts within 5g data



Brown, LCMC, Genres-Ru



Wind En vs Wind Ru over Brown categories



Conclusions

- Web collection is easy, thanks to Marco, Silvia, et al

Conclusions

- Web collection is easy, thanks to Marco, Silvia, et al
- We can identify areas of interest
- Open questions:

Conclusions

- Web collection is easy, thanks to Marco, Silvia, et al
- We can identify areas of interest
- Open questions:
 - parameters for assessing texts

Conclusions

- Web collection is easy, thanks to Marco, Silvia, et al
- We can identify areas of interest
- Open questions:
 - parameters for assessing texts
 - features for machine learning

Conclusions

- Web collection is easy, thanks to Marco, Silvia, et al
- We can identify areas of interest
- Open questions:
 - parameters for assessing texts
 - features for machine learning
 - dimensions for document similarity

New texts

