Genre Identification of Documents in a Large Web Corpus

Vít Suchomel

Natural Language Processing Centre,
Faculty of Informatics, Masaryk University

2017-03-15

Issues of Building Language Resources from the Web

Particular tasks:

- Language identification,
- Character encoding detection,
- Efficient web crawling,
- Boilerplate (unwanted content) removal,
- De-duplication (removal of identical and nearly identical texts),
- *Fighting web spam*,
- Text type: topic & *genre classification*,
- Authorship recognition,
- Storing & indexing of large text collections.

NLPC & Lexical Computing corpus tools:
`http://corpus.tools/`

Web Corpora Properties

The web is the largest corpus – ‘Web as Corpus’ (`http://sigwac.org.uk/`)

Advantages

- huge data
- various types of documents
- current form of written language
- easy to access, low cost

Disadvantages

- unordered, messy
- unwanted content (boilerplate, spam, computer generated)
- duplicates
- errors
- *What is inside?*

Definitions of Genre

“A particular style or category of works of art; esp. a type of literary work characterised by a particular form, style, or purpose.” – A genral OED definition.

“A set of conventions (regularities) that transcend individual texts, helping humans to identify the communicative purpose and the context underlying a document.” – Santini, Mehler, Sharoff: Genres on the Web: Computational Models and Empirical Studies. Vol. 42. Springer Science & Business Media, 2010.

Sketch Engine perspective: The users need to know what texts is the corpus they use based on – language research, building dictionaries, n-gram models for writing prediction,. . . The genre composition of a corpus is an important information.

Another Set of Genres?

- Incompatible genres of the BNC, the Brown-family corpora and any other studies in genre classification
- We also need to represent genres, which are specific to the Web, such as personal blogs
- A lot of disagreement between the users in assigning the genres
- We need to start with our own genre typology

Serge Sharoff's Set of 13 Genres (1 – 3)

Code	Label	Question to be answered
A1	argumentative	To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view? ('Strongly', for argumentative blogs, editorials or opinion pieces.)
A4	fictive	To what extent is the text's content fictional? ('None' if you judge it to be factual/informative.)
A7	instructive	To what extent does the text aim at teaching the reader how something works? (For example, a tutorial or an FAQ.)

Serge Sharoff's Set of 13 Genres (4 – 6)

Code	Label	Question to be answered
A8	hard news	To what extent does the text appear to be an informative report of events recent at the time of writing? (For example, a newswire. Information about future events can be hardnews too. 'None' if a news article only *discusses* a state of affairs).
A9	legal	To what extent does the text lay down a contract or specify a set of regulations? (For example, a law, a contract or copyright notices.)
A11	personal	To what extent does the text report from a first-person point of view? (For example, a diary-like blog entry.)

Serge Sharoff's Set of 13 Genres (7 – 9)

Code	Label	Question to be answered
A12	commercial	To what extent does the text promote a product or service? (For example, an advert.)
A13	ideology	To what extent is the text intended to promote a political movement, party, religious faith or other non-commercial cause? (For example, a political manifesto.)
A14	scientific /technical	To what extent would you consider the text as representing research? (For example, a research paper. Also, it can be 'Partly' if a news text reports scientific contents.)

Serge Sharoff's Set of 13 Genres (10 – 12)

Code	Label	Question to be answered
A16	informative	To what extent does the text provide information to define a topic? (For example, encyclopedic articles or text books).
A17	evaluative	To what extent does the text evaluate a specific entity by endorsing or criticising it? (For example, by providing a product review).
A20	appellative	To what extent does the text requests an action from the reader? ('Strongly' for requests, calls for papers and other appellative texts).

Serge Sharoff's Set of 13 Genres (13)

Code	**Label**	**Question to be answered**
A22	nontext	To what extent is the text different from what is expected to be a normal running text? ('Strongly' for spam, computer generated text, lists of links, online forms).

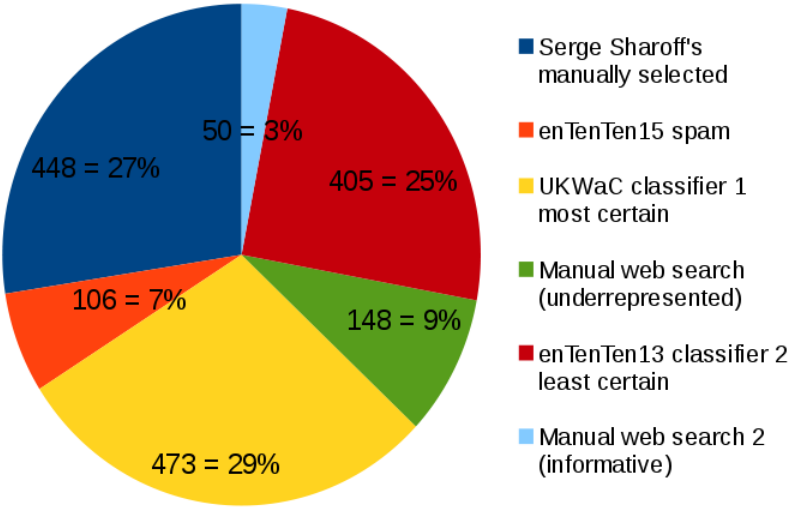
Classification Using FastText

```
fasttext supervised
fasttext predict-prob
```

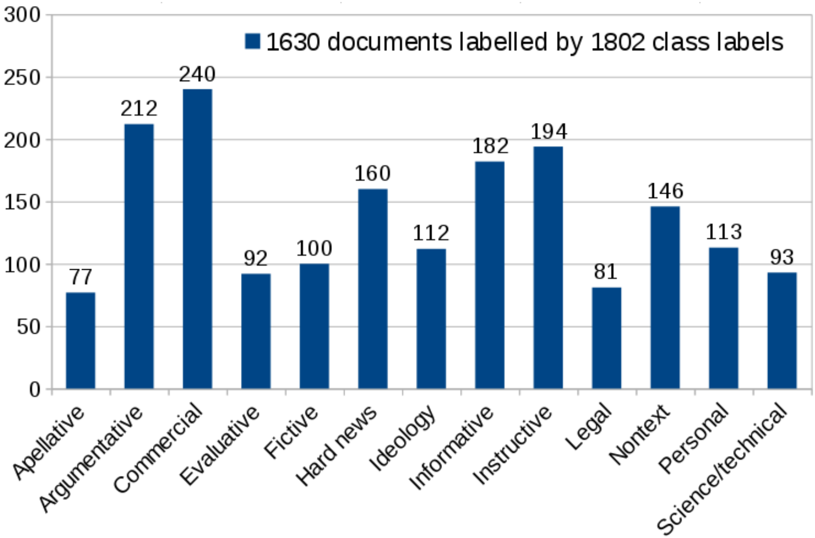
Project Workflow

- Serge's manually selected documents
- Classifier 1, evaluation
- UKWaC classifier 1 most certain documents
- enTenTen15 spam (loans, medicaments, essays, clever, other)
- Manual web search (underrepresented)
- Classifier 2, evaluation
- Active learning proof
- enTenTen13 classifier 2 least certain – active learning
- Manual web search (informative)
- Final classifier, evaluation
- Classifier applied to enTenTen15, evaluation

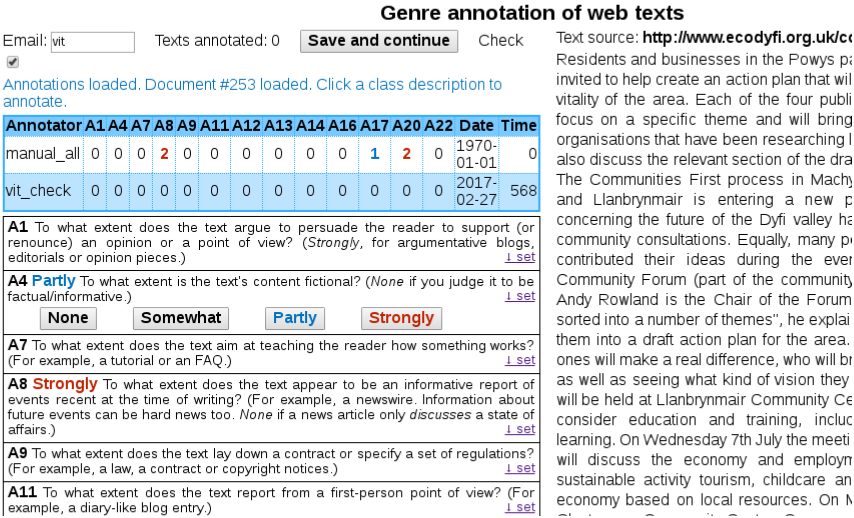
Annotated Data Collections – By Source



Annotated Data Collections – By Genre



Annotation Interface



Genre annotation of web texts

Email: vit Texts annotated: 0

☑

Annotations loaded. Document #253 loaded. Click a class description to annotate.

Annotator	A1	A4	A7	A8	A9	A11	A12	A13	A14	A16	A17	A20	A22	Date	Time
manual_all	0	0	0	2	0	0	0	0	0	0	1	2	0	1970-01-01	0
vit_check	0	0	0	0	0	0	0	0	0	0	0	0	0	2017-02-27	568

A1 To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view? (*Strongly*, for argumentative blogs, editorials or opinion pieces.) ↓ set

A4 **Partly** To what extent is the text's content fictional? (*None* if you judge it to be factual/informative.) ↓ set

A7 To what extent does the text aim at teaching the reader how something works? (For example, a tutorial or an FAQ.) ↓ set

A8 **Strongly** To what extent does the text appear to be an informative report of events recent at the time of writing? (For example, a newswire. Information about future events can be hard news too. *None* if a news article only *discusses* a state of affairs.) ↓ set

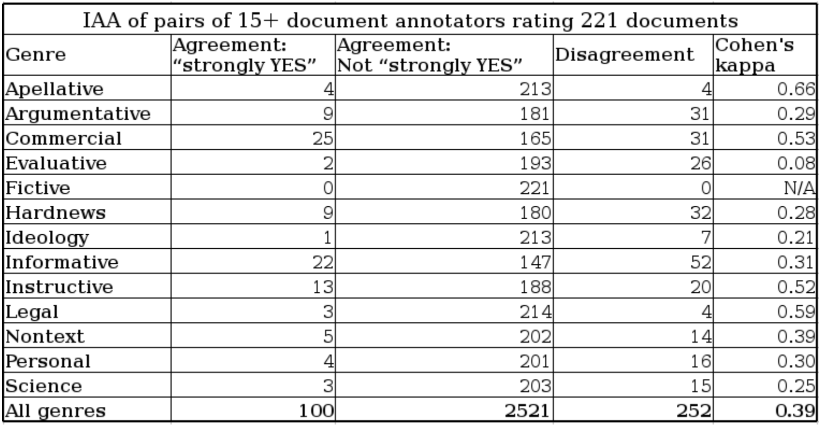
A9 To what extent does the text lay down a contract or specify a set of regulations? (For example, a law, a contract or copyright notices.) ↓ set

A11 To what extent does the text report from a first-person point of view? (For example, a diary-like blog entry.) ↓ set

Text source: **http://www.ecodyfi.org.uk/c**

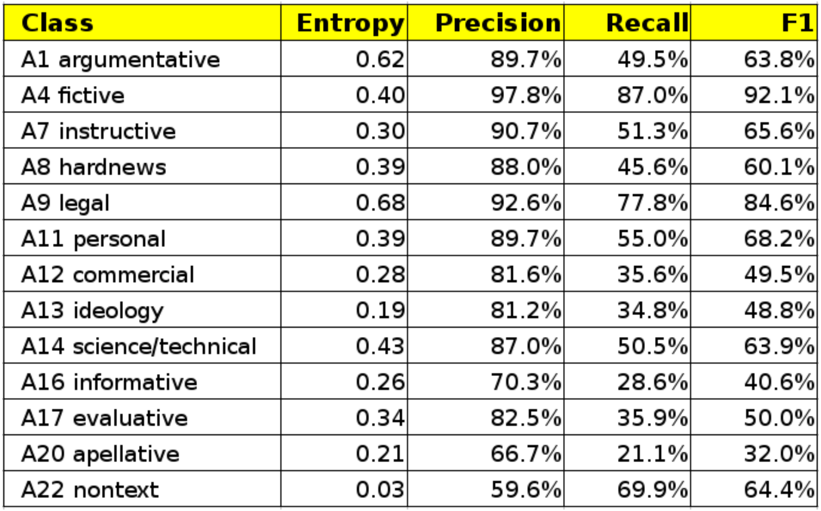
Residents and businesses in the Powys p
invited to help create an action plan that wi
vitality of the area. Each of the four publi
focus on a specific theme and will bring
organisations that have been researching
also discuss the relevant section of the dra
The Communities First process in Mach
and Llanbrynmair is entering a new p
concerning the future of the Dyfi valley h
community consultations. Equally, many p
contributed their ideas during the eve
Community Forum (part of the communit
Andy Rowland is the Chair of the Forum
sorted into a number of themes", he explai
them into a draft action plan for the area.
ones will make a real difference, who will b
as well as seeing what kind of vision they
will be held at Llanbrynmair Community C
consider education and training, inclu
learning. On Wednesday 7th July the meet
will discuss the economy and employ
sustainable activity tourism, childcare a
economy based on local resources. On

Interannotator Agreement – Not High



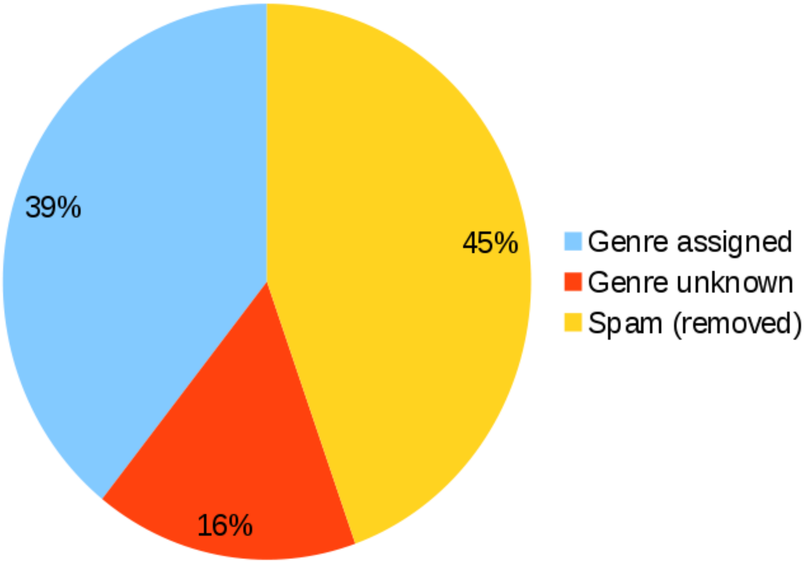
IAA of pairs of 15+ document annotators rating 221 documents				
Genre	Agreement: “strongly YES”	Agreement: Not “strongly YES”	Disagreement	Cohen's kappa
Apellative	4	213	4	0.66
Argumentative	9	181	31	0.29
Commercial	25	165	31	0.53
Evaluative	2	193	26	0.08
Fictive	0	221	0	N/A
Hardnews	9	180	32	0.28
Ideology	1	213	7	0.21
Informative	22	147	52	0.31
Instructive	13	188	20	0.52
Legal	3	214	4	0.59
Nontext	5	202	14	0.39
Personal	4	201	16	0.30
Science	3	203	15	0.25
All genres	**100**	**2521**	**252**	**0.39**

Classifier Evaluation – 30 Fold Crossvalidation

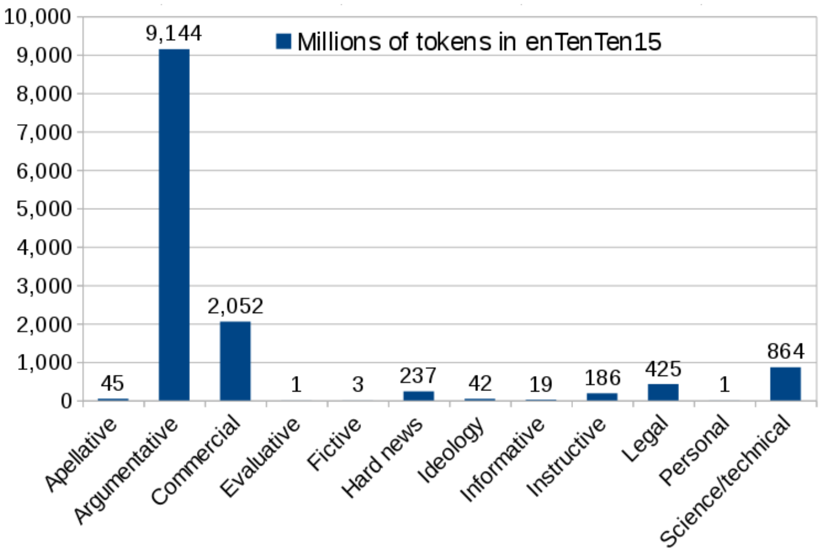


Class	Entropy	Precision	Recall	F1
A1 argumentative	0.62	89.7%	49.5%	63.8%
A4 fictive	0.40	97.8%	87.0%	92.1%
A7 instructive	0.30	90.7%	51.3%	65.6%
A8 hardnews	0.39	88.0%	45.6%	60.1%
A9 legal	0.68	92.6%	77.8%	84.6%
A11 personal	0.39	89.7%	55.0%	68.2%
A12 commercial	0.28	81.6%	35.6%	49.5%
A13 ideology	0.19	81.2%	34.8%	48.8%
A14 science/technical	0.43	87.0%	50.5%	63.9%
A16 informative	0.26	70.3%	28.6%	40.6%
A17 evaluative	0.34	82.5%	35.9%	50.0%
A20 apellative	0.21	66.7%	21.1%	32.0%
A22 nontext	0.03	59.6%	69.9%	64.4%

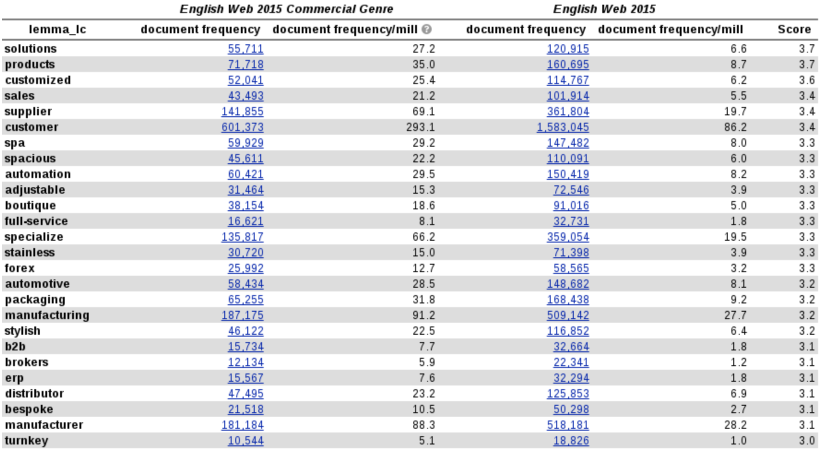
enTenTen15 Classification – Spam/Genre/Unknown



enTenTen15 Classification – Genres

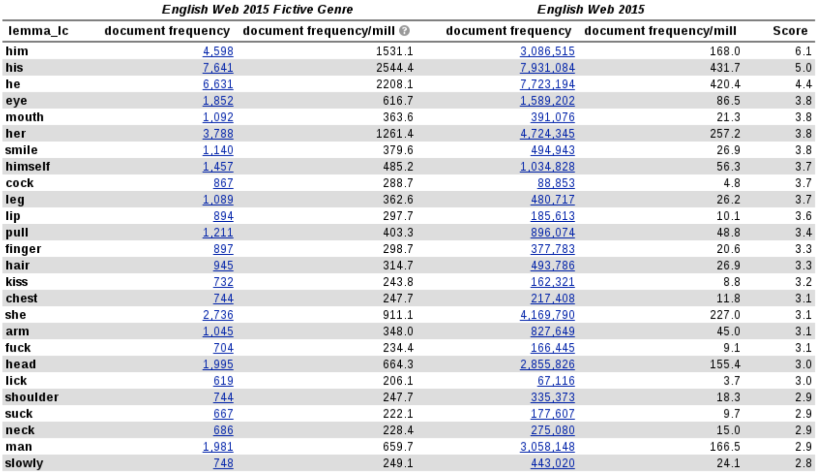


enTenTen15 Commercial vs. All – Keyword Comparison



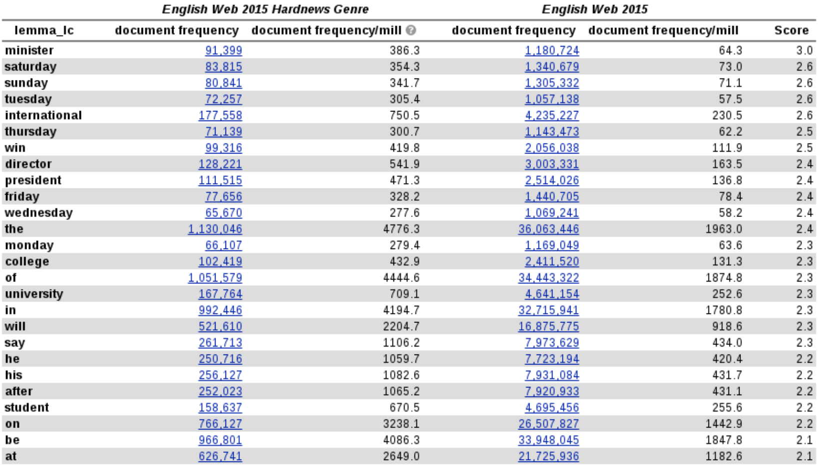
	English Web 2015 Commercial Genre		English Web 2015		
lemma_lc	document frequency	document frequency/mill	document frequency	document frequency/mill	Score
solutions	55,711	27.2	120,915	6.6	3.7
products	71,718	35.0	160,695	8.7	3.7
customized	52,041	25.4	114,767	6.2	3.6
sales	43,493	21.2	101,914	5.5	3.4
supplier	141,855	69.1	361,804	19.7	3.4
customer	601,373	293.1	1,583,045	86.2	3.4
spa	59,929	29.2	147,482	8.0	3.3
spacious	45,611	22.2	110,091	6.0	3.3
automation	60,421	29.5	150,419	8.2	3.3
adjustable	31,464	15.3	72,546	3.9	3.3
boutique	38,154	18.6	91,016	5.0	3.3
full-service	16,621	8.1	32,731	1.8	3.3
specialize	135,817	66.2	359,054	19.5	3.3
stainless	30,720	15.0	71,398	3.9	3.3
forex	25,992	12.7	58,565	3.2	3.3
automotive	58,434	28.5	148,682	8.1	3.2
packaging	65,255	31.8	168,438	9.2	3.2
manufacturing	187,175	91.2	509,142	27.7	3.2
stylish	46,122	22.5	116,852	6.4	3.2
b2b	15,734	7.7	32,664	1.8	3.1
brokers	12,134	5.9	22,341	1.2	3.1
erp	15,567	7.6	32,294	1.8	3.1
distributor	47,495	23.2	125,853	6.9	3.1
bespoke	21,518	10.5	50,298	2.7	3.1
manufacturer	181,184	88.3	518,181	28.2	3.1
turnkey	10,544	5.1	18,826	1.0	3.0

enTenTen15 Fictive vs. All – Keyword Comparison



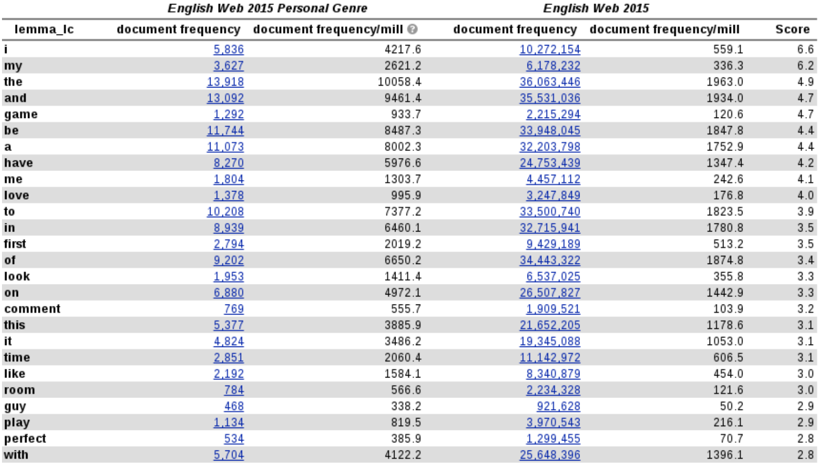
	English Web 2015 Fictive Genre		*English Web 2015*		
lemma_lc	**document frequency**	**document frequency/mill**	**document frequency**	**document frequency/mill**	**Score**
him	4,598	1531.1	3,086,515	168.0	6.1
his	7,641	2544.4	7,931,084	431.7	5.0
he	6,631	2208.1	7,723,194	420.4	4.4
eye	1,852	616.7	1,589,202	86.5	3.8
mouth	1,092	363.6	391,076	21.3	3.8
her	3,788	1261.4	4,724,345	257.2	3.8
smile	1,140	379.6	494,943	26.9	3.8
himself	1,457	485.2	1,034,828	56.3	3.7
cock	867	288.7	88,853	4.8	3.7
leg	1,089	362.6	480,717	26.2	3.7
lip	894	297.7	185,613	10.1	3.6
pull	1,211	403.3	896,074	48.8	3.4
finger	897	298.7	377,783	20.6	3.3
hair	945	314.7	493,786	26.9	3.3
kiss	732	243.8	162,321	8.8	3.2
chest	744	247.7	217,408	11.8	3.1
she	2,736	911.1	4,169,790	227.0	3.1
arm	1,045	348.0	827,649	45.0	3.1
fuck	704	234.4	166,445	9.1	3.1
head	1,995	664.3	2,855,826	155.4	3.0
lick	619	206.1	67,116	3.7	3.0
shoulder	744	247.7	335,373	18.3	2.9
suck	667	222.1	177,607	9.7	2.9
neck	686	228.4	275,080	15.0	2.9
man	1,981	659.7	3,058,148	166.5	2.9
slowly	748	249.1	443,020	24.1	2.8

enTenTen15 Hard news vs. All – Keyword Comparison



	English Web 2015 Hardnews Genre		*English Web 2015*		
lemma_lc	**document frequency**	**document frequency/mill**	**document frequency**	**document frequency/mill**	**Score**
minister	91,399	386.3	1,180,724	64.3	3.0
saturday	83,815	354.3	1,340,679	73.0	2.6
sunday	80,841	341.7	1,305,332	71.1	2.6
tuesday	72,257	305.4	1,057,138	57.5	2.6
international	177,558	750.5	4,235,227	230.5	2.6
thursday	71,139	300.7	1,143,473	62.2	2.5
win	99,316	419.8	2,056,038	111.9	2.5
director	128,221	541.9	3,003,331	163.5	2.4
president	111,515	471.3	2,514,026	136.8	2.4
friday	77,656	328.2	1,440,705	78.4	2.4
wednesday	65,670	277.6	1,069,241	58.2	2.4
the	1,130,046	4776.3	36,063,446	1963.0	2.4
monday	66,107	279.4	1,169,049	63.6	2.3
college	102,419	432.9	2,411,520	131.3	2.3
of	1,051,579	4444.6	34,443,322	1874.8	2.3
university	167,764	709.1	4,641,154	252.6	2.3
in	992,446	4194.7	32,715,941	1780.8	2.3
will	521,610	2204.7	16,875,775	918.6	2.3
say	261,713	1106.2	7,973,629	434.0	2.3
he	250,716	1059.7	7,723,194	420.4	2.2
his	256,127	1082.6	7,931,084	431.7	2.2
after	252,023	1065.2	7,920,933	431.1	2.2
student	158,637	670.5	4,695,456	255.6	2.2
on	766,127	3238.1	26,507,827	1442.9	2.2
be	966,801	4086.3	33,948,045	1847.8	2.1
at	626,741	2649.0	21,725,936	1182.6	2.1

enTenTen15 Personal vs. All – Keyword Comparison



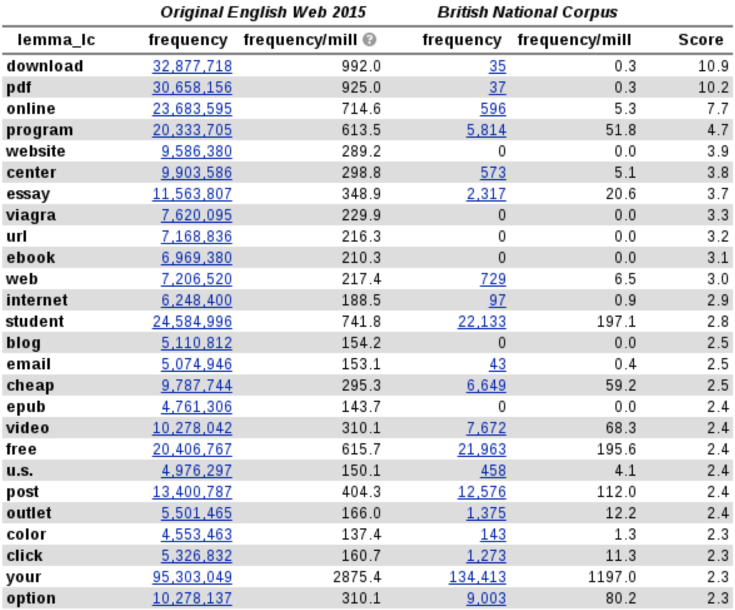
	English Web 2015 Personal Genre		*English Web 2015*		
lemma_lc	**document frequency**	**document frequency/mill**	**document frequency**	**document frequency/mill**	**Score**
i	5,836	4217.6	10,272,154	559.1	6.6
my	3,627	2621.2	6,178,232	336.3	6.2
the	13,918	10058.4	36,063,446	1963.0	4.9
and	13,092	9461.4	35,531,036	1934.0	4.7
game	1,292	933.7	2,215,294	120.6	4.7
be	11,744	8487.3	33,948,045	1847.8	4.4
a	11,073	8002.3	32,203,798	1752.9	4.4
have	8,270	5976.6	24,753,439	1347.4	4.2
me	1,804	1303.7	4,457,112	242.6	4.1
love	1,378	995.9	3,247,849	176.8	4.0
to	10,208	7377.2	33,500,740	1823.5	3.9
in	8,939	6460.1	32,715,941	1780.8	3.5
first	2,794	2019.2	9,429,189	513.2	3.5
of	9,202	6650.2	34,443,322	1874.8	3.4
look	1,953	1411.4	6,537,025	355.8	3.3
on	6,880	4972.1	26,507,827	1442.9	3.3
comment	769	555.7	1,909,521	103.9	3.2
this	5,377	3885.9	21,652,205	1178.6	3.1
it	4,824	3486.2	19,345,088	1053.0	3.1
time	2,851	2060.4	11,142,972	606.5	3.1
like	2,192	1584.1	8,340,879	454.0	3.0
room	784	566.6	2,234,328	121.6	3.0
guy	468	338.2	921,628	50.2	2.9
play	1,134	819.5	3,970,543	216.1	2.9
perfect	534	385.9	1,299,455	70.7	2.8
with	5,704	4122.2	25,648,396	1396.1	2.8

Removing Spam from enTenTen15 – Indicative Words

Corpus sizes and relative frequencies (number of occurrences per million words) of selected words in the original enTenTen15 compared to the same corpus without documents classified as spam:

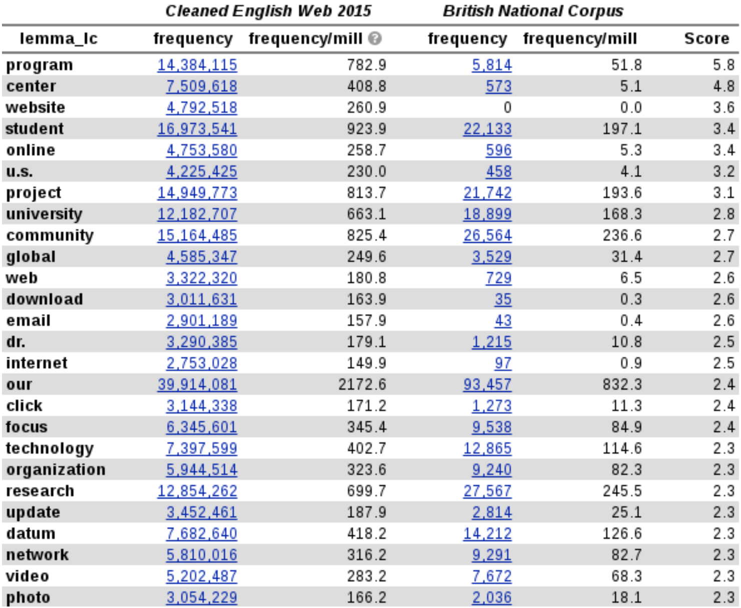
Count	Original	Spam removed	Kept
Corpus size (documents)	58,438,034	37,810,139	65 %
Corpus size (tokens)	33,144,241,513	18,371,812,861	55 %
“viagra”	229.70	3.42	1 %
“cialis 20 mg”	2.70	0.02	1 %
“aspirin”	5.60	1.50	15 %
“loan”	166.30	48.34	29 %
“payday loan”	24.20	1.10	5 %
“cheap”	295.30	64.30	22 %
“essay”	348.90	33.95	5 %
“essay writing”	26.60	0.57	1 %
“pass the exam”	0.34	0.36	59 %

Original enTenTen15 vs. BNC – Keyword Comparison



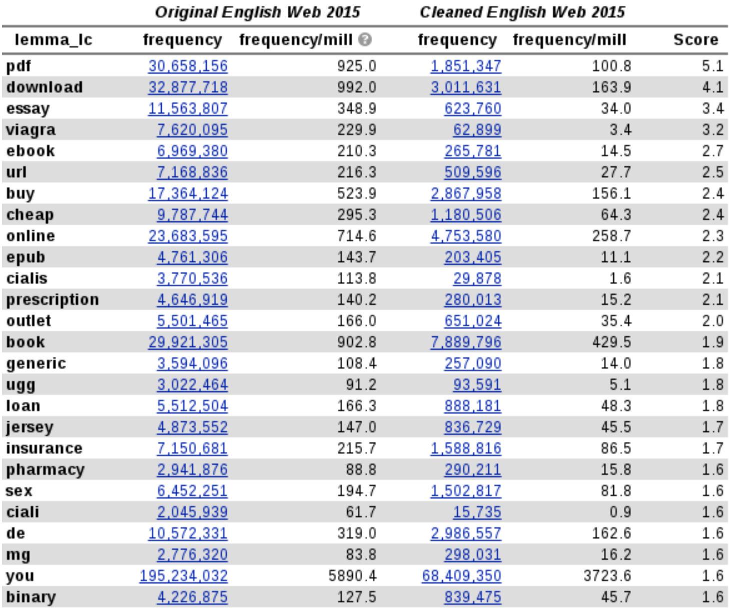
	Original English Web 2015		British National Corpus		
lemma_lc	frequency	frequency/mill	frequency	frequency/mill	Score
download	32,877,718	992.0	35	0.3	10.9
pdf	30,658,156	925.0	37	0.3	10.2
online	23,683,595	714.6	596	5.3	7.7
program	20,333,705	613.5	5,814	51.8	4.7
website	9,586,380	289.2	0	0.0	3.9
center	9,903,586	298.8	573	5.1	3.8
essay	11,563,807	348.9	2,317	20.6	3.7
viagra	7,620,095	229.9	0	0.0	3.3
url	7,168,836	216.3	0	0.0	3.2
ebook	6,969,380	210.3	0	0.0	3.1
web	7,206,520	217.4	729	6.5	3.0
internet	6,248,400	188.5	97	0.9	2.9
student	24,584,996	741.8	22,133	197.1	2.8
blog	5,110,812	154.2	0	0.0	2.5
email	5,074,946	153.1	43	0.4	2.5
cheap	9,787,744	295.3	6,649	59.2	2.5
epub	4,761,306	143.7	0	0.0	2.4
video	10,278,042	310.1	7,672	68.3	2.4
free	20,406,767	615.7	21,963	195.6	2.4
u.s.	4,976,297	150.1	458	4.1	2.4
post	13,400,787	404.3	12,576	112.0	2.4
outlet	5,501,465	166.0	1,375	12.2	2.4
color	4,553,463	137.4	143	1.3	2.3
click	5,326,832	160.7	1,273	11.3	2.3
your	95,303,049	2875.4	134,413	1197.0	2.3
option	10,278,137	310.1	9,003	80.2	2.3

Cleaned enTenTen15 vs. BNC – Keyword Comparison



	Cleaned English Web 2015		British National Corpus		
lemma_lc	frequency	frequency/mill	frequency	frequency/mill	Score
program	14,384,115	782.9	5,814	51.8	5.8
center	7,509,618	408.8	573	5.1	4.8
website	4,792,518	260.9	0	0.0	3.6
student	16,973,541	923.9	22,133	197.1	3.4
online	4,753,580	258.7	596	5.3	3.4
u.s.	4,225,425	230.0	458	4.1	3.2
project	14,949,773	813.7	21,742	193.6	3.1
university	12,182,707	663.1	18,899	168.3	2.8
community	15,164,485	825.4	26,564	236.6	2.7
global	4,585,347	249.6	3,529	31.4	2.7
web	3,322,320	180.8	729	6.5	2.6
download	3,011,631	163.9	35	0.3	2.6
email	2,901,189	157.9	43	0.4	2.6
dr.	3,290,385	179.1	1,215	10.8	2.5
internet	2,753,028	149.9	97	0.9	2.5
our	39,914,081	2172.6	93,457	832.3	2.4
click	3,144,338	171.2	1,273	11.3	2.4
focus	6,345,601	345.4	9,538	84.9	2.4
technology	7,397,599	402.7	12,865	114.6	2.3
organization	5,944,514	323.6	9,240	82.3	2.3
research	12,854,262	699.7	27,567	245.5	2.3
update	3,452,461	187.9	2,814	25.1	2.3
datum	7,682,640	418.2	14,212	126.6	2.3
network	5,810,016	316.2	9,291	82.7	2.3
video	5,202,487	283.2	7,672	68.3	2.3
photo	3,054,229	166.2	2,036	18.1	2.3

Original vs. Cleaned enTenTen15 – Keyword Comparison



	Original English Web 2015		*Cleaned English Web 2015*		
lemma_lc	frequency	frequency/mill	frequency	frequency/mill	Score
pdf	30,658,156	925.0	1,851,347	100.8	5.1
download	32,877,718	992.0	3,011,631	163.9	4.1
essay	11,563,807	348.9	623,760	34.0	3.4
viagra	7,620,095	229.9	62,899	3.4	3.2
ebook	6,969,380	210.3	265,781	14.5	2.7
url	7,168,836	216.3	509,596	27.7	2.5
buy	17,364,124	523.9	2,867,958	156.1	2.4
cheap	9,787,744	295.3	1,180,506	64.3	2.4
online	23,683,595	714.6	4,753,580	258.7	2.3
epub	4,761,306	143.7	203,405	11.1	2.2
cialis	3,770,536	113.8	29,878	1.6	2.1
prescription	4,646,919	140.2	280,013	15.2	2.1
outlet	5,501,465	166.0	651,024	35.4	2.0
book	29,921,305	902.8	7,889,796	429.5	1.9
generic	3,594,096	108.4	257,090	14.0	1.8
ugg	3,022,464	91.2	93,591	5.1	1.8
loan	5,512,504	166.3	888,181	48.3	1.8
jersey	4,873,552	147.0	836,729	45.5	1.7
insurance	7,150,681	215.7	1,588,816	86.5	1.7
pharmacy	2,941,876	88.8	290,211	15.8	1.6
sex	6,452,251	194.7	1,502,817	81.8	1.6
ciali	2,045,939	61.7	15,735	0.9	1.6
de	10,572,331	319.0	2,986,557	162.6	1.6
mg	2,776,320	83.8	298,031	16.2	1.6
you	195,234,032	5890.4	68,409,350	3723.6	1.6
binary	4,226,875	127.5	839,475	45.7	1.6

Conclusion

- Genre classifier
- Working active learning scheme (we proved active learning helps for this more than selecting random documents)
- Separate thresholds for genres favouring precision over recall
- enTenTen15 annotated, more corpora to follow
- Non-text based classifier helps identifying spam
- Future work: Are genre features preserved by machine translation of texts?

Photo: Kathleen & Ryan Rush