Dr Gabriele Salciute Civiliene

gabriele.salciute-civiliene@kcl.ac.uk

Coding & the Humanities

Week 6 | Part 1_**Unstructured Data**

16/11/2020

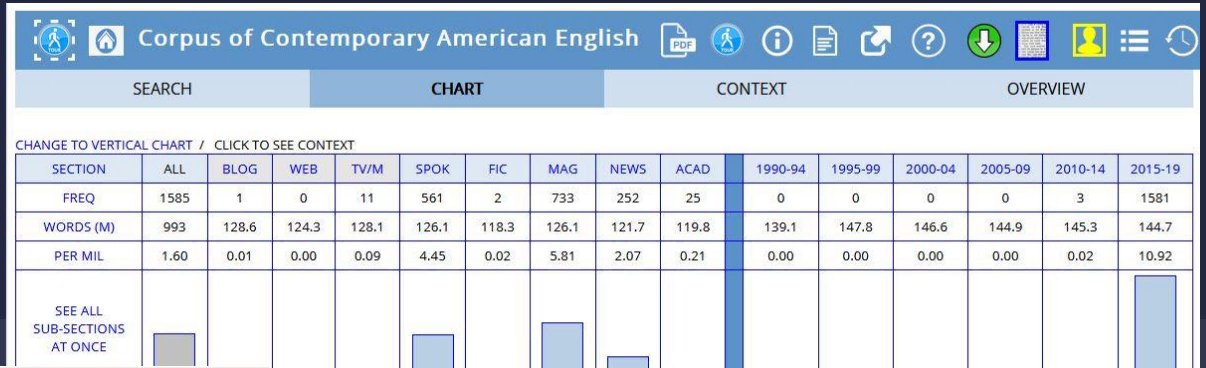
In this part …

- You’ll learn about the principles of analysing natural language data which is often termed as unstructured data
- Natural language data can be collected and stored as texts, audio, video or web data files

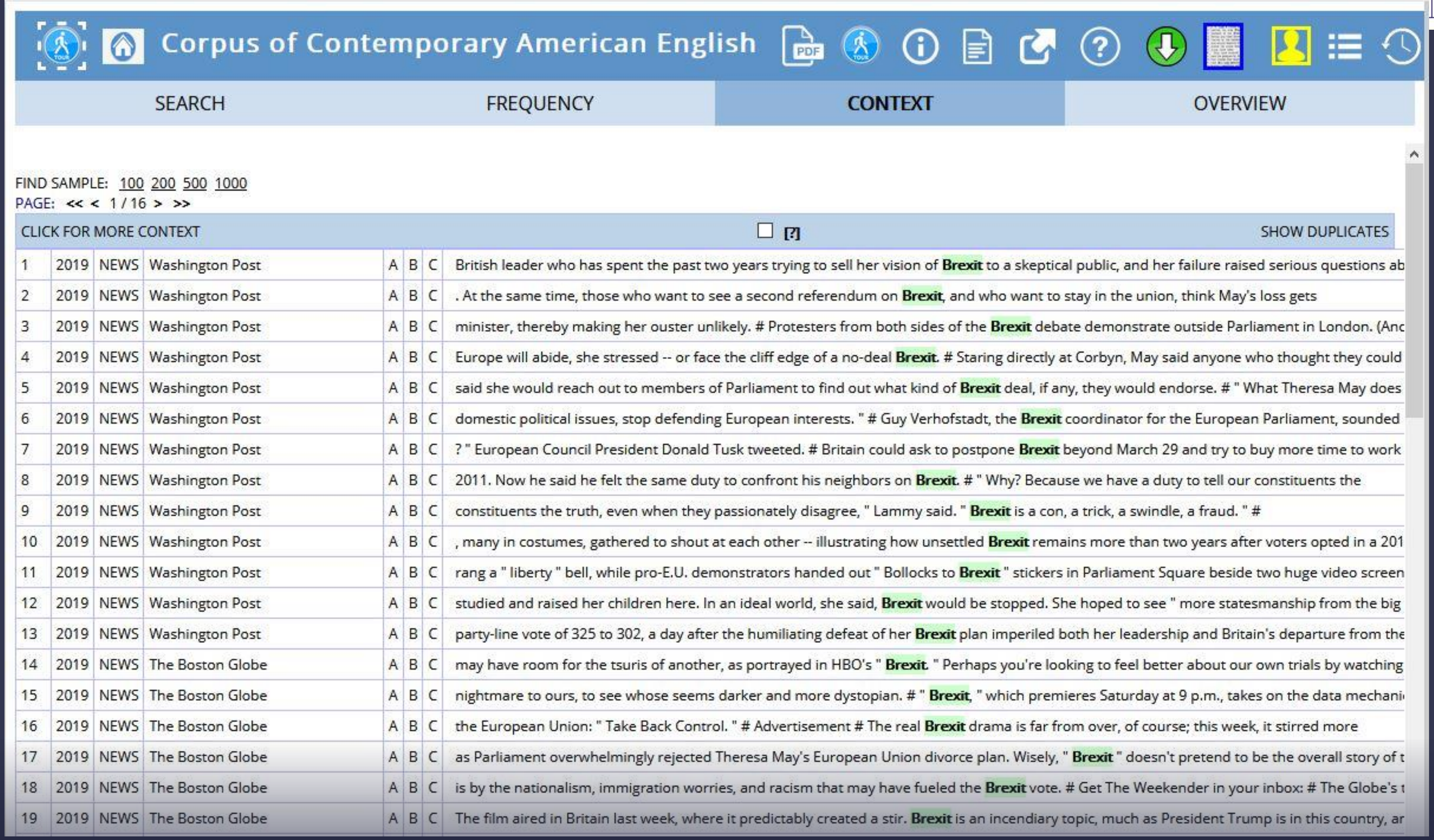
What is a corpus

- A corpus (pl. corpora) is a collection of texts that have been structured, annotated and organized for machine languages to query and extract natural language data
- Different types of corpora, e.g. parallel, monolingual, diachronic, etc
- Corpus building is an expensive and labour-intensive task that not every language or variety affords
- NLP approach, though it follows corpus techniques, does not depend on building a corpus

COCA (Corpus of Contemporary American English)
https://www.english-corpora.org/coca/


Corpus of Contemporary American English
SEARCH
CHART
CONTEXT
OVERVIEW
CHANGE TO VERTICAL CHART / CLICK TO SEE CONTEXT

SECTION	ALL	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD		1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
FREQ	1585	1	0	11	561	2	733	252	25		0	0	0	0	3	1581
WORDS (M)	993	128.6	124.3	128.1	126.1	118.3	126.1	121.7	119.8		139.1	147.8	146.6	144.9	145.3	144.7
PER MIL	1.60	0.01	0.00	0.09	4.45	0.02	5.81	2.07	0.21		0.00	0.00	0.00	0.00	0.02	10.92
SEE ALL SUB-SECTIONS AT ONCE																

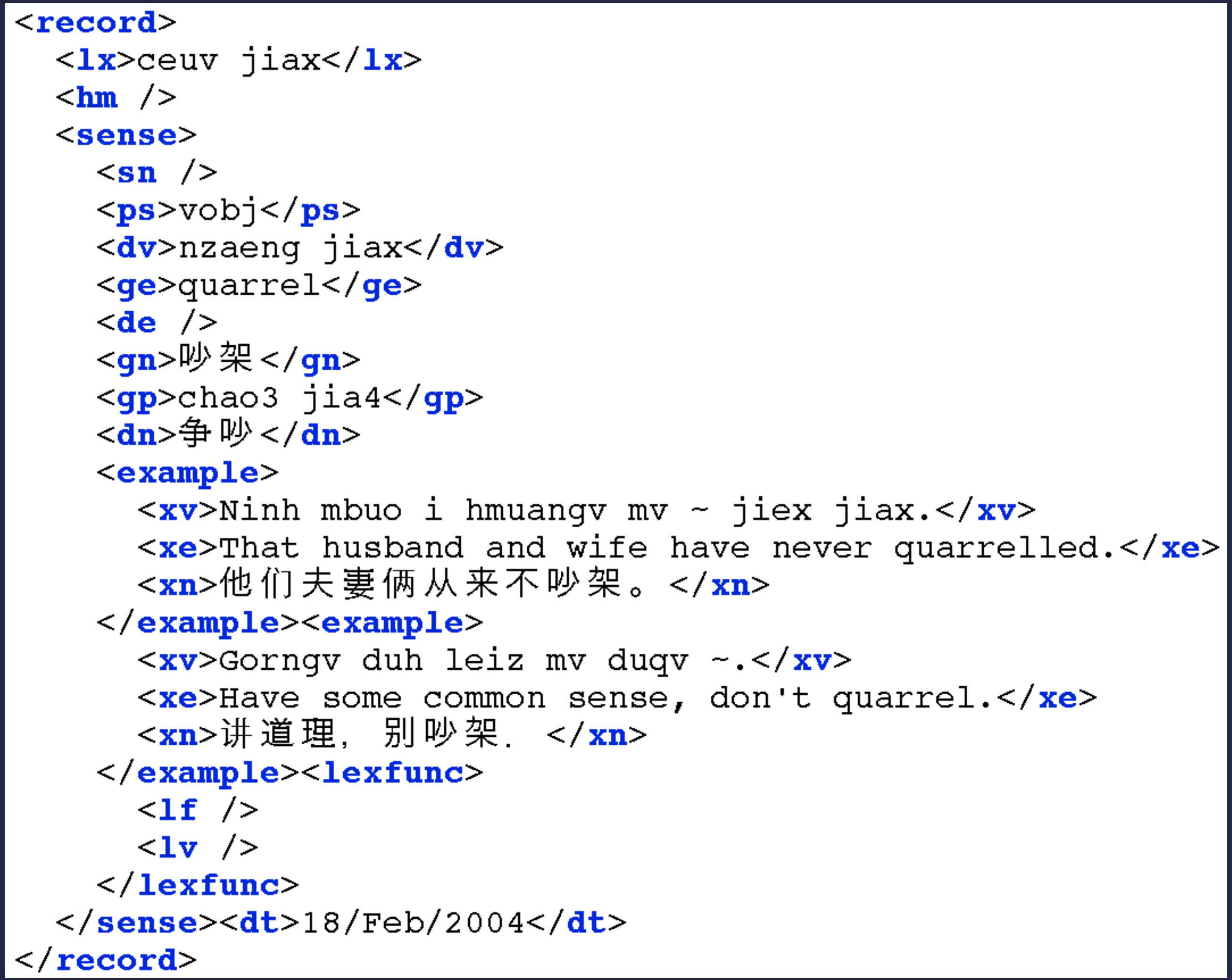


FIND SAMPLE: 100 200 500 1000
PAGE: << < 1 / 16 > >>

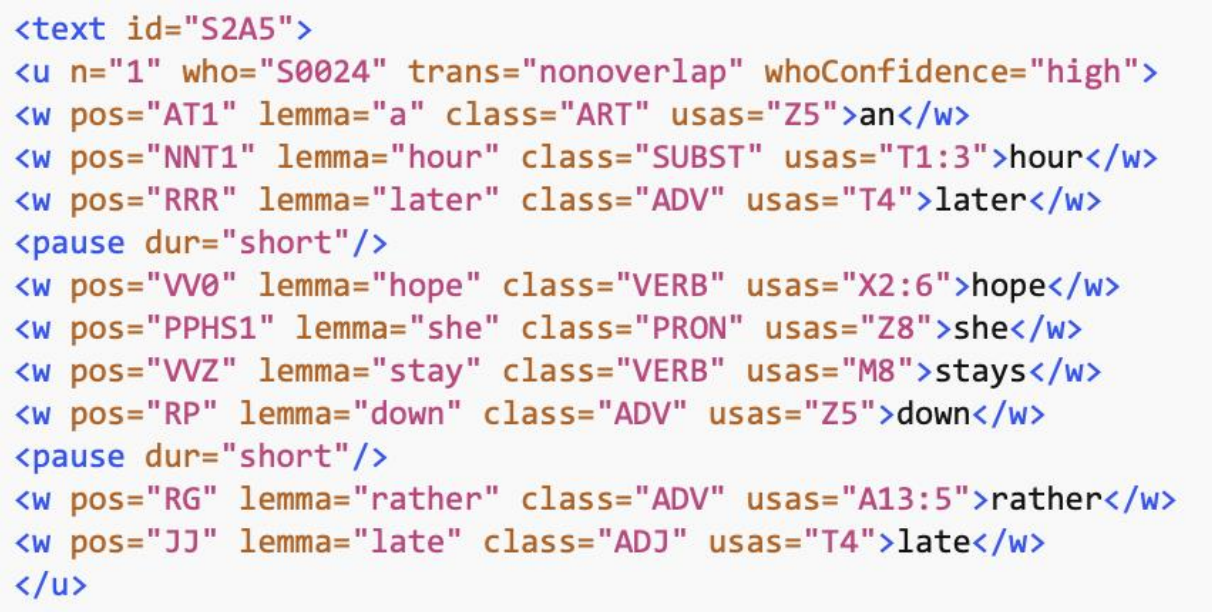
CLICK FOR MORE CONTEXT □ [?] SHOW DUPLICATES

1	2019	NEWS	Washington Post	A	B	C	British leader who has spent the past two years trying to sell her vision of **Brexit** to a skeptical public, and her failure raised serious questions ab
2	2019	NEWS	Washington Post	A	B	C	. At the same time, those who want to see a second referendum on **Brexit**, and who want to stay in the union, think May's loss gets
3	2019	NEWS	Washington Post	A	B	C	minister, thereby making her ouster unlikely. # Protesters from both sides of the **Brexit** debate demonstrate outside Parliament in London. (Anc
4	2019	NEWS	Washington Post	A	B	C	Europe will abide, she stressed -- or face the cliff edge of a no-deal **Brexit**. # Staring directly at Corbyn, May said anyone who thought they could
5	2019	NEWS	Washington Post	A	B	C	said she would reach out to members of Parliament to find out what kind of **Brexit** deal, if any, they would endorse. # " What Theresa May does
6	2019	NEWS	Washington Post	A	B	C	domestic political issues, stop defending European interests. " # Guy Verhofstadt, the **Brexit** coordinator for the European Parliament, sounded
7	2019	NEWS	Washington Post	A	B	C	? " European Council President Donald Tusk tweeted. # Britain could ask to postpone **Brexit** beyond March 29 and try to buy more time to work
8	2019	NEWS	Washington Post	A	B	C	2011. Now he said he felt the same duty to confront his neighbors on **Brexit**. # " Why? Because we have a duty to tell our constituents the
9	2019	NEWS	Washington Post	A	B	C	constituents the truth, even when they passionately disagree, " Lammy said. " **Brexit** is a con, a trick, a swindle, a fraud. " #
10	2019	NEWS	Washington Post	A	B	C	, many in costumes, gathered to shout at each other -- illustrating how unsettled **Brexit** remains more than two years after voters opted in a 201
11	2019	NEWS	Washington Post	A	B	C	rang a " liberty " bell, while pro-E.U. demonstrators handed out " Bollocks to **Brexit** " stickers in Parliament Square beside two huge video screen
12	2019	NEWS	Washington Post	A	B	C	studied and raised her children here. In an ideal world, she said, **Brexit** would be stopped. She hoped to see " more statesmanship from the big
13	2019	NEWS	Washington Post	A	B	C	party-line vote of 325 to 302, a day after the humiliating defeat of her **Brexit** plan imperiled both her leadership and Britain's departure from the
14	2019	NEWS	The Boston Globe	A	B	C	may have room for the tsuris of another, as portrayed in HBO's " **Brexit**. " Perhaps you're looking to feel better about our own trials by watching
15	2019	NEWS	The Boston Globe	A	B	C	nightmare to ours, to see whose seems darker and more dystopian. # " **Brexit**, " which premieres Saturday at 9 p.m., takes on the data mechani
16	2019	NEWS	The Boston Globe	A	B	C	the European Union: " Take Back Control. " # Advertisement # The real **Brexit** drama is far from over, of course; this week, it stirred more
17	2019	NEWS	The Boston Globe	A	B	C	as Parliament overwhelmingly rejected Theresa May's European Union divorce plan. Wisely, " **Brexit** " doesn't pretend to be the overall story of t
18	2019	NEWS	The Boston Globe	A	B	C	is by the nationalism, immigration worries, and racism that may have fueled the **Brexit** vote. # Get The Weekender in your inbox: # The Globe's t
19	2019	NEWS	The Boston Globe	A	B	C	The film aired in Britain last week, where it predictably created a stir. **Brexit** is an incendiary topic, much as President Trump is in this country, ar

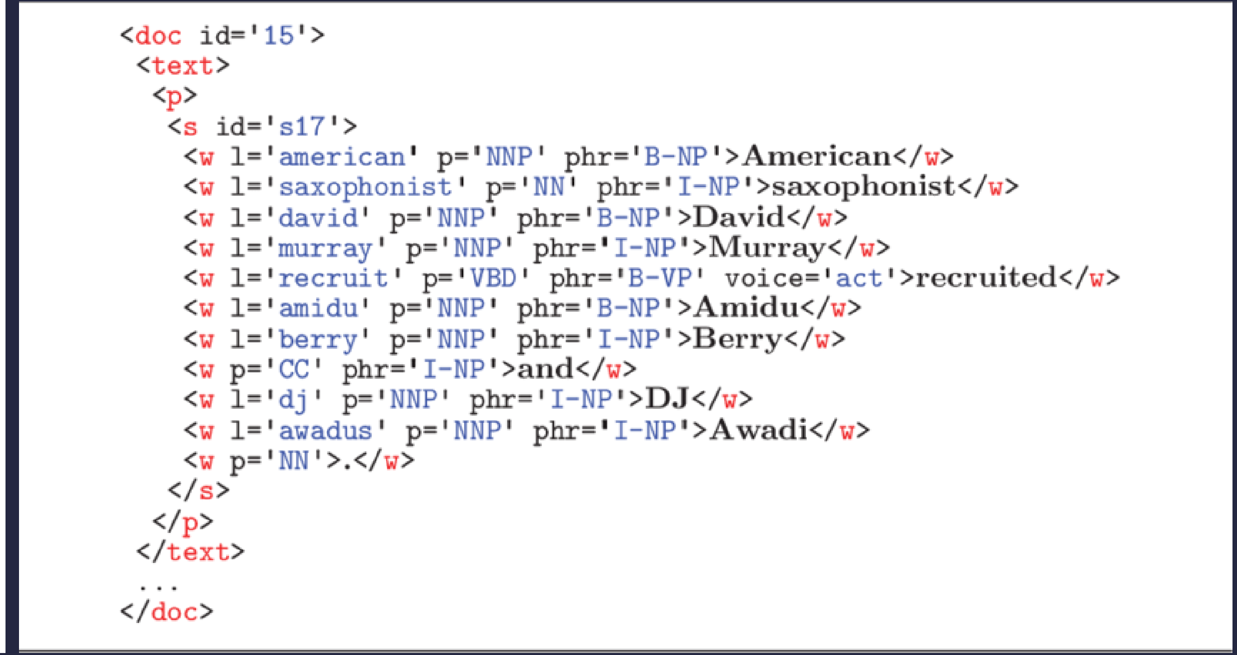
XML-structured texts



```
<record>
  <lx>ceuv jiax</lx>
  <hm />
  <sense>
    <sn />
    <ps>vobj</ps>
    <dv>nzaeng jiax</dv>
    <ge>quarrel</ge>
    <de />
    <gn>吵架</gn>
    <gp>chao3 jia4</gp>
    <dn>争吵</dn>
    <example>
      <xv>Ninh mbuo i hmuangv mv ~ jiex jiax.</xv>
      <xe>That husband and wife have never quarrelled.</xe>
      <xn>他们夫妻俩从来不吵架。</xn>
    </example><example>
      <xv>Gorngv duh leiz mv duqv ~.</xv>
      <xe>Have some common sense, don't quarrel.</xe>
      <xn>讲道理，别吵架．</xn>
    </example><lexfunc>
      <lf />
      <lv />
    </lexfunc>
  </sense><dt>18/Feb/2004</dt>
</record>
```



```
<text id="S2A5">
<u n="1" who="S0024" trans="nonoverlap" whoConfidence="high">
<w pos="AT1" lemma="a" class="ART" usas="Z5">an</w>
<w pos="NNT1" lemma="hour" class="SUBST" usas="T1:3">hour</w>
<w pos="RRR" lemma="later" class="ADV" usas="T4">later</w>
<pause dur="short"/>
<w pos="VV0" lemma="hope" class="VERB" usas="X2:6">hope</w>
<w pos="PPHS1" lemma="she" class="PRON" usas="Z8">she</w>
<w pos="VVZ" lemma="stay" class="VERB" usas="M8">stays</w>
<w pos="RP" lemma="down" class="ADV" usas="Z5">down</w>
<pause dur="short"/>
<w pos="RG" lemma="rather" class="ADV" usas="A13:5">rather</w>
<w pos="JJ" lemma="late" class="ADJ" usas="T4">late</w>
</u>
```

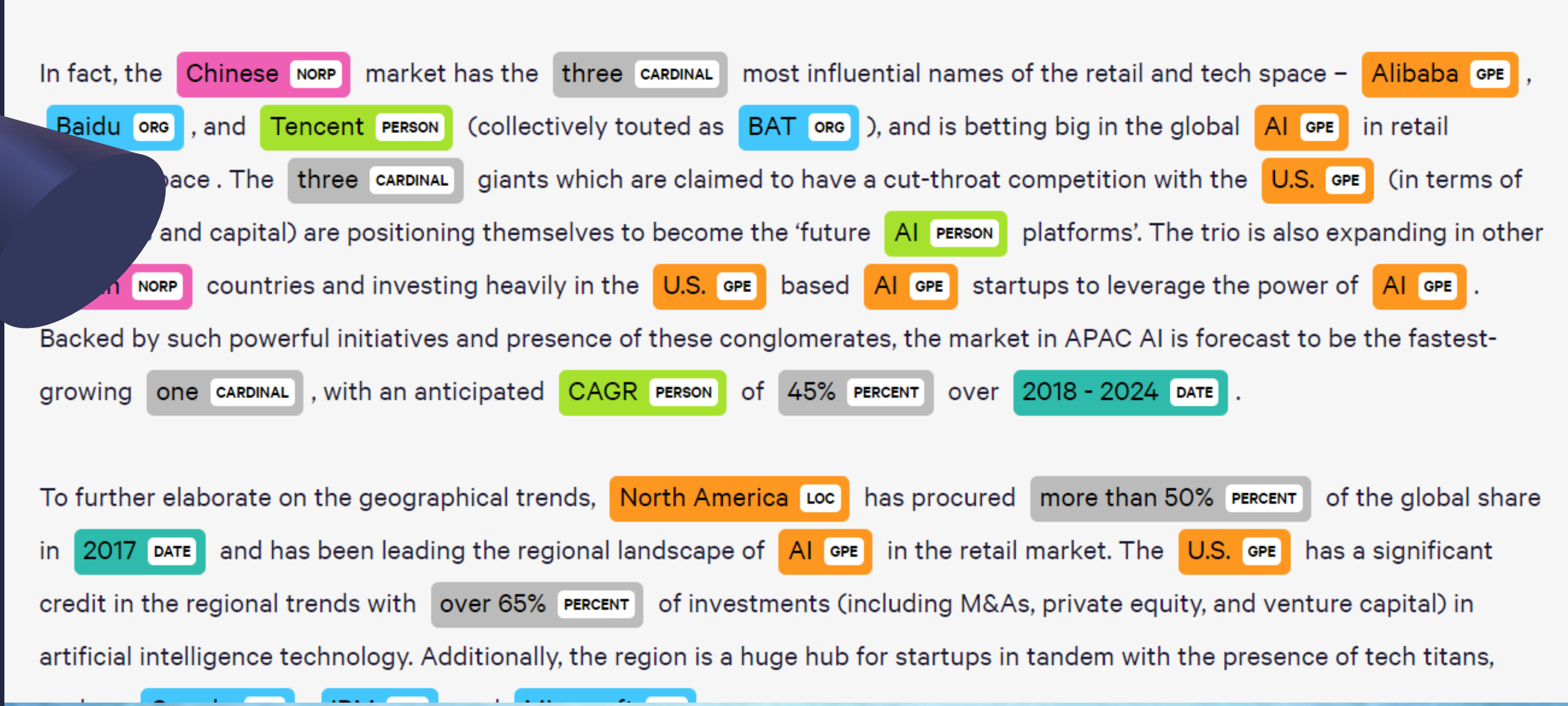


```
<doc id='15'>
 <text>
  <p>
   <s id='s17'>
    <w l='american' p='NNP' phr='B-NP'>American</w>
    <w l='saxophonist' p='NN' phr='I-NP'>saxophonist</w>
    <w l='david' p='NNP' phr='B-NP'>David</w>
    <w l='murray' p='NNP' phr='I-NP'>Murray</w>
    <w l='recruit' p='VBD' phr='B-VP' voice='act'>recruited</w>
    <w l='amidu' p='NNP' phr='B-NP'>Amidu</w>
    <w l='berry' p='NNP' phr='I-NP'>Berry</w>
    <w p='CC' phr='I-NP'>and</w>
    <w l='dj' p='NNP' phr='I-NP'>DJ</w>
    <w l='awadus' p='NNP' phr='I-NP'>Awadi</w>
    <w p='NN'>.</w>
   </s>
  </p>
 </text>
 ...
</doc>
```

Types of extractable corpus & NL data

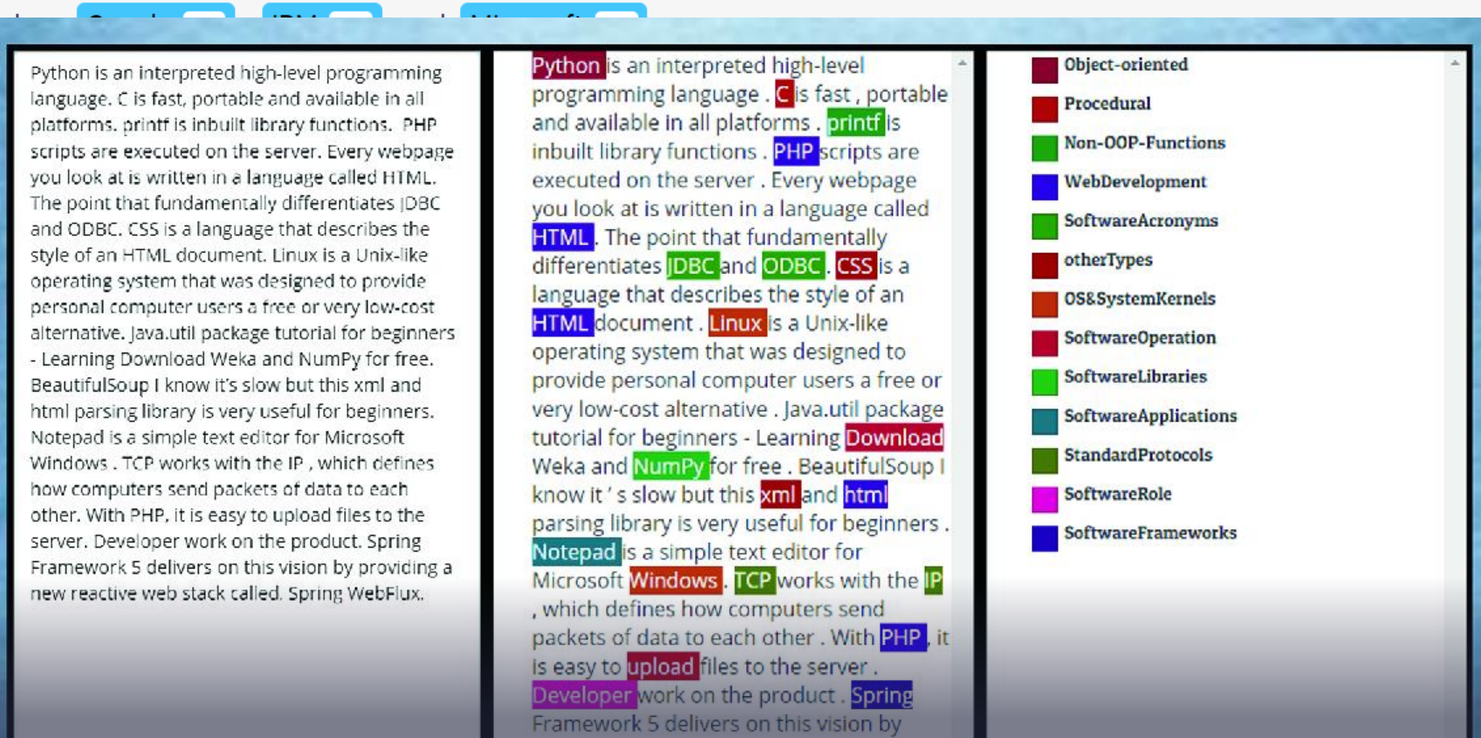
- The type of data that we can extract from texts include word and sentence frequencies, n-grams (e.g. collocations), word groups by POS (parts-of-speech) taggers, semantic patterns, named entities, & the like.

Named entities



In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail pace . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other n NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans,

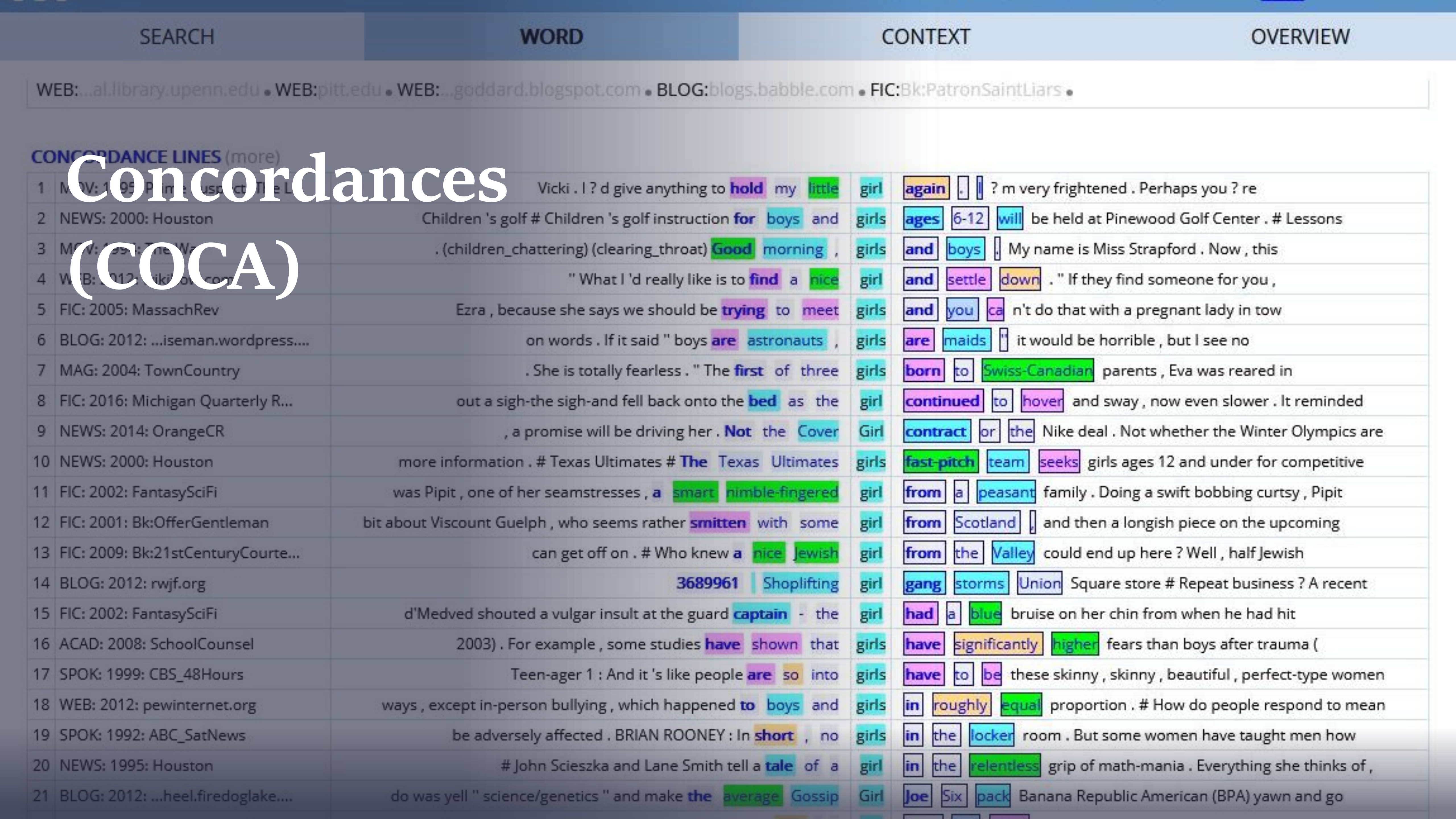


Python is an interpreted high-level programming language. C is fast, portable and available in all platforms. printf is inbuilt library functions. PHP scripts are executed on the server. Every webpage you look at is written in a language called HTML. The point that fundamentally differentiates JDBC and ODBC. CSS is a language that describes the style of an HTML document. Linux is a Unix-like operating system that was designed to provide personal computer users a free or very low-cost alternative. Java.util package tutorial for beginners - Learning Download Weka and NumPy for free. BeautifulSoup I know it's slow but this xml and html parsing library is very useful for beginners. Notepad is a simple text editor for Microsoft Windows . TCP works with the IP , which defines how computers send packets of data to each other. With PHP, it is easy to upload files to the server. Developer work on the product. Spring Framework 5 delivers on this vision by providing a new reactive web stack called. Spring WebFlux.

Python is an interpreted high-level programming language . C is fast , portable and available in all platforms . printf is inbuilt library functions . PHP scripts are executed on the server . Every webpage you look at is written in a language called HTML . The point that fundamentally differentiates JDBC and ODBC . CSS is a language that describes the style of an HTML document . Linux is a Unix-like operating system that was designed to provide personal computer users a free or very low-cost alternative . Java.util package tutorial for beginners - Learning Download Weka and NumPy for free . BeautifulSoup I know it ' s slow but this xml and html parsing library is very useful for beginners . Notepad is a simple text editor for Microsoft Windows . TCP works with the IP , which defines how computers send packets of data to each other . With PHP , it is easy to upload files to the server . Developer work on the product . Spring Framework 5 delivers on this vision by

🟥 Object-oriented
🟥 Procedural
🟩 Non-OOP-Functions
🟦 WebDevelopment
🟩 SoftwareAcronyms
🟥 otherTypes
🟥 OS&SystemKernels
🟥 SoftwareOperation
🟩 SoftwareLibraries
🟦 SoftwareApplications
🟩 StandardProtocols
🟪 SoftwareRole
🟦 SoftwareFrameworks

Concordances (COCA)



WEB:...al.library.upenn.edu • WEB:pitt.edu • WEB:...goddard.blogspot.com • BLOG:blogs.babble.com • FIC:Bk:PatronSaintLiars •

CONCORDANCE LINES (more)

1	MOV: 1995: Prime Suspect: The L...	Vicki . I ? d give anything to hold my little	girl	again . I ? m very frightened . Perhaps you ? re
2	NEWS: 2000: Houston	Children 's golf # Children 's golf instruction for boys and	girls	ages 6-12 will be held at Pinewood Golf Center . # Lessons
3	MOV: 1994: The Wa...	. (children_chattering) (clearing_throat) Good morning ,	girls	and boys . My name is Miss Strapford . Now , this
4	WEB: 2012: wikihow.com	" What I 'd really like is to find a nice	girl	and settle down . " If they find someone for you ,
5	FIC: 2005: MassachRev	Ezra , because she says we should be trying to meet	girls	and you ca n't do that with a pregnant lady in tow
6	BLOG: 2012: ...iseman.wordpress....	on words . If it said " boys are astronauts ,	girls	are maids " it would be horrible , but I see no
7	MAG: 2004: TownCountry	. She is totally fearless . " The first of three	girls	born to Swiss-Canadian parents , Eva was reared in
8	FIC: 2016: Michigan Quarterly R...	out a sigh-the sigh-and fell back onto the bed as the	girl	continued to hover and sway , now even slower . It reminded
9	NEWS: 2014: OrangeCR	, a promise will be driving her . Not the Cover	Girl	contract or the Nike deal . Not whether the Winter Olympics are
10	NEWS: 2000: Houston	more information . # Texas Ultimates # The Texas Ultimates	girls	fast-pitch team seeks girls ages 12 and under for competitive
11	FIC: 2002: FantasySciFi	was Pipit , one of her seamstresses , a smart nimble-fingered	girl	from a peasant family . Doing a swift bobbing curtsy , Pipit
12	FIC: 2001: Bk:OfferGentleman	bit about Viscount Guelph , who seems rather smitten with some	girl	from Scotland , and then a longish piece on the upcoming
13	FIC: 2009: Bk:21stCenturyCourte...	can get off on . # Who knew a nice Jewish	girl	from the Valley could end up here ? Well , half Jewish
14	BLOG: 2012: rwjf.org	3689961 | Shoplifting	girl	gang storms Union Square store # Repeat business ? A recent
15	FIC: 2002: FantasySciFi	d'Medved shouted a vulgar insult at the guard captain - the	girl	had a blue bruise on her chin from when he had hit
16	ACAD: 2008: SchoolCounsel	2003) . For example , some studies have shown that	girls	have significantly higher fears than boys after trauma (
17	SPOK: 1999: CBS_48Hours	Teen-ager 1 : And it 's like people are so into	girls	have to be these skinny , skinny , beautiful , perfect-type women
18	WEB: 2012: pewinternet.org	ways , except in-person bullying , which happened to boys and	girls	in roughly equal proportion . # How do people respond to mean
19	SPOK: 1992: ABC_SatNews	be adversely affected . BRIAN ROONEY : In short , no	girls	in the locker room . But some women have taught men how
20	NEWS: 1995: Houston	# John Scieszka and Lane Smith tell a tale of a	girl	in the relentless grip of math-mania . Everything she thinks of ,
21	BLOG: 2012: ...heel.firedoglake....	do was yell " science/genetics " and make the average Gossip	Girl	Joe Six pack Banana Republic American (BPA) yawn and go

What is NLKT?

- NLTK stands for ‘Natural language toolkit’
- It’s a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning
- It also contains over 50 corpora and lexical resources (http://www.nltk.org/nltk_data/) necessary to carry out the tasks of tokenizing, tagging and otherwise structuring texts