

Course Topics

- Text Analysis analysing unstructured data
- **W** Python
- Regular Expressions
- Natural Language Toolkit (NLTK)



It is agreat blessing and happiness to a natio when the Hing imployeth such a man as you are to chet and do for him who I'm personadid his the are and fear of God on Sum. job was a just man and aperfect and the cause that he know not he fearehed out to deliver The poor and oppressed and him that had none to he Riem, affakerne for on in your office. I have the Honor to be your Relation and I know you have much interest with Lord Greange if you can make Peace for of loveing my husband to much, he knowed very we that he was my job and now god his made him a vode lo Scourgeth me. * * * * * * * * *

"inuch fuller account then this and he work it down you may from





Course Structure

Anticipate about ~7 hours/week

- 2 hour course meeting, 2-4 Wednesday
- 1 assignment per week, ~2 hours
- Office hours on request
- Independent learning, ~2 hours
- Teams for introductions, meetings, office hours, questions, files

gr Hilda Jan: 20 173 It is agreat blessing and happiness to a natio when the Hing imployeth such a man as you are to chet and do for him who I'm personadid his the are and fear of God on Sum. job was a just man and aperfect and the the poor and oppressed and him that had none to he liem, affakerne for on in your office. I have the Honor to be your Relation and I know you have much interest with Lord Greange if you can make Peace for nie you know the promices that is to the Dear make of loveing my husband to much, he knowes very we that he was my jobol and now god his made him a vode lo Scourgeth me. * * * * * * * * *



INTRODUCTIONS

- Why are you interested in text analysis?
- Have you used Python before?
- Have you used Jupyter Notebooks before?
- Have you used Regular Expressions before?
- Have you used NLTK before?



questions



Participant Expectations

Wednesday classes are introductions to material
Assignments will be given on Thursday
Classes are not recorded but all class materials will
be uploaded to Teams
Please let me know in advance if you cannot attend!
Message me on Teams to schedule office hours for

82 Hilda Jan: 20 173 It is agreat blessing and happiness to a natio when the Hing imployeth such a man as you are to chet and do for him who I'm personadid his the are and fear of God on him. job was a just man and a perfect and the cause that he know not he searched out to deliver the poor and oppressed and him that had none to he hiem, affallerne for on in your office. I have the Monor to be your Relation and I know you have much interest with Lord Greasure if you can make Peace for nie you know the promices that is to the Dear make of loveing my husband to much, he knowed very we that he was my job and now god his made him a vode lo Scourgeth me. * * * * * * * * *





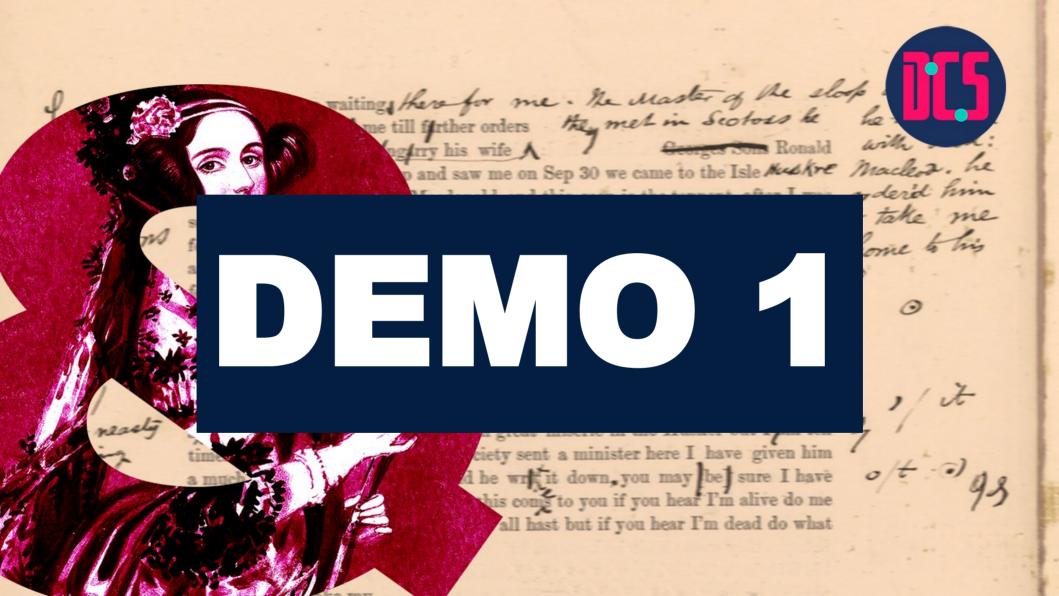
Course Software

Jupyter Notebooks / Jupyterlabs

- With Notable https://www.ed.ac.uk/information-services/learning-technology/noteable/accessing-noteable
 - After logging into MyEd: https://noteable.edina.ac.uk/launch
- With Google Colab https://colab.research.google.com
- Locally (install with pip/pip3 or conda)











Further Resources

- Noteable User Guide:
 https://noteable.edina.ac.uk/user_guide/#hide_ge_7
- Jupyter Notebooks in Noteable: https://github.com/edina/Exemplars2020/blob/master/TeachingDocs/T utorials/UsingNoteableBeginner.ipynb
- Jupyter Notebooks: https://glam-workbench.github.io/getting-started/
- Python: https://programminghistorian.org/en/lessons/introductionand-installation









NLTK

Natural Language Toolkit

Natural language = human language
= "unstructured" data









NLTK

Examples of data sources for natural language:

- Books
- Newspapers
- Magazines
- Websites
- Transcriptions of audio (i.e. interview, movie dialogue)
- Social media

Always read the licensing/copyright information and terms of use!





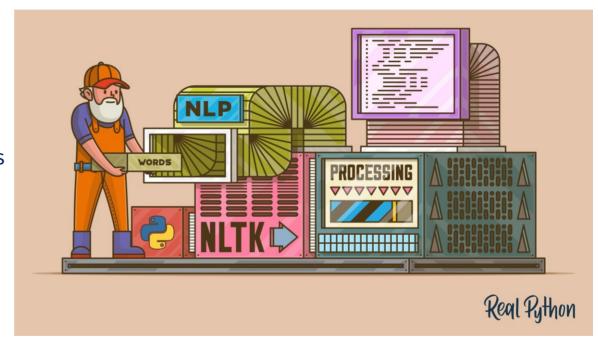




Why use NLTK

What kinds of questions can you ask when you can use a programming language to study hundreds, thousands, or even millions of pages of digital text?

"Distant reading"









Why use NLTK

What kinds of questions can you ask when you can use a programming language to study hundreds, thousands, or even millions of pages of digital text?

"Distant reading"

NLTK Isn't everything

What kinds of questions can you ask when you can physically hold and look at a printed text, be it an original publication or later edition of the text?

"Close reading"
Book history







NLTK Terminology

Tokens vs. words

Digitized vs. digital

Normalization (a.k.a. standardization)

Document vs. corpus vs. corpora









Summarising a Text

Built-in functions include:

len(text)

sorted(vocabulary of text)

NLTK Text methods include:

Text.count(word)

Apocalyplis Blo.ozdi.

(8 a Oficibre saib. Olio fidelib a sgiles x leuce funt in pos motione boson oper. b Alenie. Ib a opandorin cade fi-de no ouerrao bereies feitness (if mall. v ed oe pits' midd oi neries. c Ex cogr. ex . q. o. Doubt o oberie e e lelliciti y at banc cend y cenir y officiare qui file by negamentri viaboble file.

cabents:qrmulta pate nocere vobis en ifi auxilio dei lupare poteritio. dimi milli fat. Antippo et lui apil majores penas patier de alij ad filina/ dinčilit q viu coburit. c Et cere ri.i.fegee cop miore pena punientur. velilli q erut ceteri.i.ad ch querii oc eifimado: vinent beo. f Et omnes ques,i.očo feti velecrati füt pe pea illo rū t be pfeetu frateli. En.XX E vidi angela. Dieta vanatioe

b babylonie z anticht z přendo apřou řispponit bánationě ípě biaboliu řípponit cám čre z ípě d bá-nabří z če šlů potezit vincerca. Ali d nabíří a de ilii poterit vinecté-Alii q in bálil adež no potit tilla ligare čříti voluti glosticat^eniče čino oefiruce potitit. D žibyti. Edyt Temboslo coda liptor vet ije obabořája oba šer niz opnititi stefernat. 1 žippben dit. apeccaffe obdit ob še imuně a poč lite obabořocetátí. k Ziganit čít. Solědá gy tilr i abesa c aliej náchbas p Selekin g illr i abssa i alış indetibus vlabol'inte ligar'iki in iline pitib'iled in illiş liganıt ipcə farari ipi, in illiş ipe ips adurnik liganıt. 1. Ermisi ibe ejecet'a fidelib'i malio cepi təfis ri arroct'. B legik in enangeliceş intra att in porces. [8] Signanit

Mico. delvra

et buobomilibopedira 5 boftes quon claffe venerant animolius predentes pftranerut ex boftib marimam multi tudinë u relidut conerli fiit ad fugam maine trenamila magna pare a niig bul retëra. vid magna pare a niig ful retëra. vië § tobes: Et vidi vnum angeli. 1. přiarebá bierofolymitanus. — Steré in fole. peninětla vite do-trineršni illá Lect. 1. Bič fol refulgés ficitle refulit in replo pei. b Et cla. e Dices ofb'ani. rc. qap pgreffa ad

elº erbonationé factă ques fatis babu. erat ad comedenda. o Er vidt be ftia.i.egypti foldana. e Et reges terre.abl inbiectos. f Et egercit" co.co.in mari i in terra. g Ad faci endü plifi ve.i.ch rege balduino et ep ercitu fue. b Et appbela e beftia. s ve.i.rex egypti v ppbete falli qui in a/ fturia fina fuerut coprebell.na orifi el? go cogitanerat enenit. 1 Et & fecit

a f in lumine intelligerie feripruran t fernose

lum ftate in foler damauit vo ce magna bices:oib amb g vo ct teplabant, a f p carbolica fide q cois &. b f .t.

labat p medin celi. Weiter ogre a f .i.ad ipm bei d voben refectio erre. f vt ber lecremini i cruciare ipiop ra maiop @ minora. gamini ad cena magna vei: vt 1 maducetif carnes regu a carnel tribunon t carnes fortili t car/ nes cora z fedetia in ipis: z car

nes oim liberozű et fuop a pu/ fillozum et magnop. Et vidi be :

ftia z reacs terre z exercitufeon u # .i.corra gpm 2 fuee blafebande 7 mbulan

ogregatos ad faciedű pliű:cům illo d fedebat i equo z cu eperci/

cũ illa pleudo ppheta: z qui fe/ i cit ligna coză ipo quib? ledurit k

eos qui acceperüt characterem « γ τι udat dei ci gaudio fotopien. h f γ lot beftie τ q adotauerüt imagine ! ci°. Ülihi milli füt bi vuo in fta/ m †

gnű ignis ardétis t fulphuris. et ceteri occifi füt in gladio fede . tie sup equi qui predit de ote f ifi". Et oce aues saturate sunt £a. XX carnibus cozum.

Mico. delvra

guine petorie et veferibit ifta laus Gith ad ecelefia trifipba tem a quantu ad militante à pebet quâth pot triupbati cons formari. fm go bein ille pfar Sugne matris gandin repfenformar, migo vil nia giurappin in a para più di control de la control de

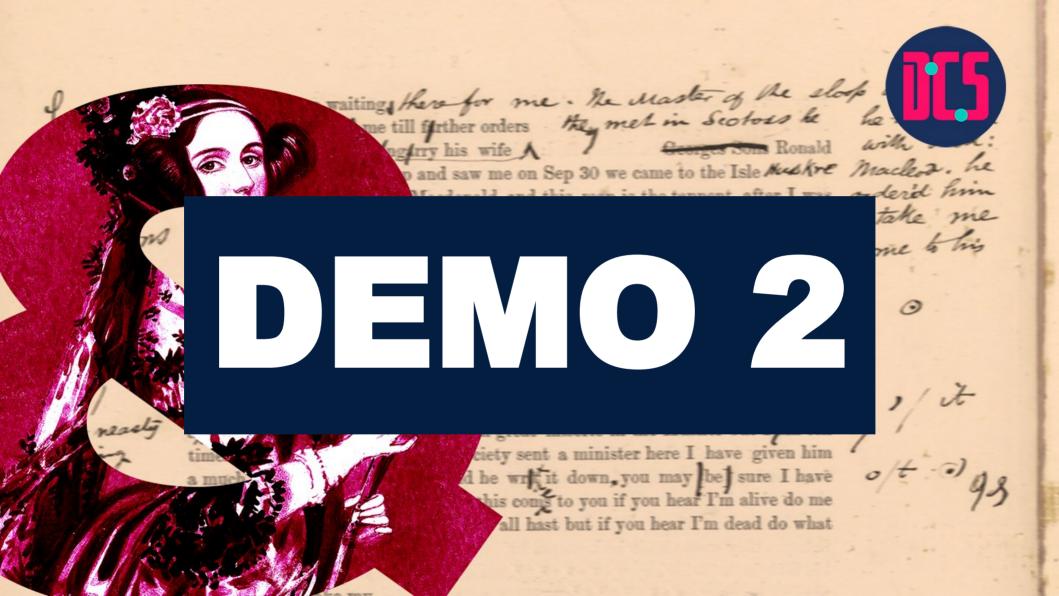
ht Ercecidi ad pedes ciote, Boffe, ditur or non folium bebet gratia redde peo pe fuio baficno: fed etis fuio mini firie p quoe reselat bolbus verirates vel alia bifficia fubminufrat. † Et birit mibi vide ne fecens te.per B ar, mune boice superbiqui querunt ide bire ab aligs bonozari. nt 200 bee vidi ccium apertum 42.boc fraliter in telligi videt be rpo per es que fequin ur. TDorslurr aurpot erponi be qiv bet et" beno vicario in fpualibus cuiufmodi funt prelati: vel tpalib" cuiul modi funt principes terreni t con fub. din boni qui parati funt ets obedire i licinis a boncitie do fubditur: ot Er exercitus qui funt in celo fede bantur cum in equis albis tellin cor

posibus mundis. by Etipe calent torcular tell, bonns phi vicari punit inftemalefacta. † Et babet in ver ftimento e femore ino ieriptum rep regum 42. qu finb fe plures bă regentes babet qui reges bici poffunt a regedo de gar boc feripta in veftimento a in fe more fuo.femur eft intrinfect boirve. ftimenta aut extrinfect. In actiboaute intrinfecte talis fupiozis appet q por puta fibi fubdită vult picto mo gaber cie be q dr. v. Etb. q nem beiper nem ftans in fole eft pdicator conflanter pe dicas 3 obstinatos feueritate inflicie vinincideo inbdit: c† De aubus 1.Demonibus & funt executores biaine i.bemonibus quant excutores sunne inflicie in shantis, quant paces fignizione formation e tradit paces fignizione formation e Et maduz cette carmes regis 42.1. obdinatorii in culpa io fequifi: b * Et apphenia e befisa e c.f.ecf.; m † 2Duffi iunt in flazione de comparatorii e de com gnum ignis ardentio. Ca.XX p E vidi angelā. Defepta in go e nerali s ipēali feptima plagas 5

tem De celo: babentes cla/ q qui occalisto papa, a benrico impato

Reference: https://www.nltk.org/book/ch01.html

Biblia Sacra by N/A - University of Edinburgh, United Kingdom - CC BY. https://www.europeana.eu/item/9200261/BibliographicResource 300005 8482943







Getting to know a text

NLTK Text methods include:

Text.concordance("word",

lines=20)

Text.similar("word")

Text.common contexts(["list",

"of", "words"])

Text.dispersion plot(["list",

"of", "words"])

Apocalypsis La. XX Blo.ozdi.

S a Offilhus sulb" (Oliv fidelih") agilies viences funt in puo moniene bonou porg. b Alenite. Dis opandonie cade h de no ouerfao berefes fesquese fir malla. Ved se pith' middioi nerile. c Ex cogre ex e. e. Dulliñ oebeite e lelliciti vi ad banc cena oenie y collision qui tilb ti pagmanerit'i babolo fino

chbetts:qemultă pat'e nocere vobis că tă auxilio dei lupare poteritis. di iliai milli fat. Antipo et lui apli maiores penas patiet de alijad filitu/ dine illie' q viu coburit. e Et cete ri.i.fegees con miore pena punientur. eifimador vinent Deo. f Etomnes ques,i.očo feti velectati füt ve pea illo ru t be pfeetu fraten. En.XX E vidi angeth. Dicta vanaride

b babylonie a antirbi a pfendo apton inponit pănation îpe piabelia pomit câm die a tipa d păr nabifia de illu poterit vincere filli d nabíří a de ilii poterit vinecté-Alii q in bálil adež no potit tilla ligare čříti voluti glosticat^eniče čino oefiruce potitit. D žibyti. Edyt Temboslo coda liptor vet ije obabořája oba šer niz opnititi stefernat. 1 žippben dit. apeccaffe obdit ob še imuně a poč lite obabořocetátí. k Ziganit čít. Solědá gy tilr i abesa c aliej náchbas p piabol fint ligar lic in iftio phtib ifed in illis liganit (peo futuri rpi. in illis tpe ros aduentel ligauit. I Et milit eu.qe eiect"a fidelib" I malio cepit ofia ri atroct*.b legit in enangelio: e intra uit in porceo. (§ Gignanit

Mico. delvra

et buobemilib pedira 5 boftes quon claffe venerant animoffus pecdentes pftranerut ex boftibomaximam multi tudine a relidui conerfi füt ad fugam maine i friidintonern un ad rugm de claffe etnamilla magna pare a niig fult retëta.dië ĝi tobeo: Et vidi vnum angeli.dipfiarebă bierofolymitaniz. Stete in folege entiella vice dov etrineren illa Lect.l. Bië fol refulgës ficitle refulit in replo pei. b Et cla.

e Dices ofb'ani. rc.qap pgreffa ad elº erbonationé factă ques fatis babu. erat ad comedenda. o Er vidt be ftia.i.covoti foldana. e Etreges terre.abi inbiecros. f Et erercit co.co.in mari i in terra. g Ad faci endü plifi ve.i.ch rege balduino et ep ercitu fue. b Et appbela e beftia. s ve.i.rex egypti v ppbete falli qui in a/ fturia fina fuerut coprebell.na orifi el? go cogitanerat enenit. 1 Et & fecit

a f in lumine ferelligerie feripruran t fernose lum ftate in foler damauit vo a fily comerfano i erlie eft. b f.i.celeftia con ce magna vices:oib auib q vo et eiplabant. a f p carbolica fide q cois &. b f .t.

labat p medin celi. Eleiter ogre a f .i.ad ipm bei d voben refectio erre. f vt ber lecremini i cruciare iprop ra maiop @ minora. gamini ad cena magna Dei: vt 1 maducetif carnes requia carnel tribunoz z carnes fortili z car/ nes cora z fedetia in ipis: z car

nes olm liberozű et fuop a pu/ fillozum et magnop. Et vidi be a

ftia z reacs terre z exercitufeon u # .i.corra gpm 2 fuee blafebande 7 mbulan ogregatos ad faciedű pliű:cům s illo d ledebati equo z cu exerci/

a f .i. cormaliter punira. tu ci? . Et apprebela e beftia: et be nari. of Et vi.vnu ange.ftanzein fole.p fole pot intelligi elaritao infili cu illa pleudo ppheta: z qui fe/ i citfigna coză ipo quib" fedurit h

cos qui acceperiit characterem » f vriudei dei ci gaudio fotopien. h f 2fot beftie t d adorauerut imagine i ei". Bihi miffi füt bi ouo in fta, mt

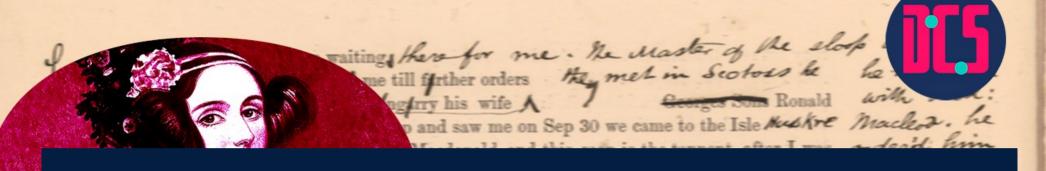
6 pprer cencheag. gnű ignis ardétis t fulpburis. et ceteri occifi fut in gladio fede . tie lug equi qui predit de ote f ibi". Et oce aues faturate funt carnibus corum. La. XX

Mico. delvra guine petorie et veferibit ifta laus Gith ad ecelefia trifipba tem a quantu ad militante à pebet quâth pot triupbati cons formari.fm qu' bein ille pfar Sugne matrie gandin repfenformar, m go vien un principin mai par de la partir bugo-nis de fetó victore de la frego ado alugo autem bugo-nis de fetó victore do la frego ado alugo autem bugo-nis de fetó victore do la frego ado alugo autem bugo-nis de fetó victore do la frego ado a frego al terra discrite allelula, p militante.

kt Ercecidi ad pedeo erte.p Bofter ditur or non folium bebet gratia redde peo pe fuio bafieno: fed etis fuio mini ffrie p quoe reselat boibas veritates vel alia bifficia fubminufrat. † Et birit mibi vide ne fecens te.per B ar, guant boles superbi qui querunt ide bite ab alije bonozari. n† 13º bee vidi celum apertum se.boc fraliter in telligt videt be roo per es que lequin tur. TDoraliter aut pot exponi be div iufmodi funt peclati:vel tpalib" cuiuf modi funt principes terreni t con fub. din boni qui parati funt ets obedire i licinis a boncitie do fubditur: ot Et exercitue qui funt in celo fede

bantur eum in equis albis 42.i.in con posibus mundis. by Etipe calent torcular tell, bonns phi vicari punit inftemalefacta. † Et babet in ver ftimento e femore ino ieriptum rep regum 42. qu finb fe plures bă regentes babet qui reges bici poffunt a regedo de gar boc feripta in veftimento a in fe more fuo.femur eft intrinfect boirve ftimenta aut ertrinfecu. In actiboaute intrinfecte talis fupiozis appet q por pulà fibi fubdità vult picto mo gaber cie be q dr. v. Etb. o nem beiper nem lucifer eit ita namirabilis. Engel igif ftans in fole eft pdicator conflanter pe dicas 3 obstinatos feueritate inflicie vinincideo inbdit: c† De aubus 1-bemonibue à finne executores biuine i.bemonibus quant excutores sunne inflicie in shantis, quant paces fignizione formation e tradit paces fignizione formation e Et maduz cette carmes regis 42.1. obdinatorii in culpa io fequifi: b * Et apphenia e befisa e c.f.ecf.; m † 2Duffi iunt in flazione de comparatorii e de com gnum ignis ardentio. Ca.XX p E vidi angelā. Defepta in go e nerali s ipēali feptima plagas 5

par veferibit ecelle pmotio no partem a fedam in parte exponent ali tem De celo: babentes cla/ q qui De caligno papa. a benrico impato



DEMO 3-5

time to the ciety sent a minister here I have given him a much a much the writing the writ

0/t . 0) 93





The Building Blocks

- Tokenization words/punctuation, sentences
- Normalization
- Stemming and lemmatizing
- Frequency counts
- Part-of-speech tagging









Tokenisation

- Tokenisation involves breaking down a piece of text into smaller units called tokens.
- Tokens can be individual words, sentences, or even characters, depending on the level of granularity desired.
- Tokenisation helps in standardizing and organizing text data, making it easier to analyse and process.
- Word-based tokenisation breaks down text into individual words, treating each word as a separate token.





Text cleaning & pre-processing

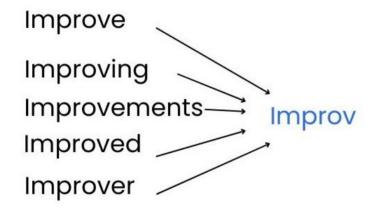
- Formatting text for analysis and removing extraneous informationWorkflows vary depending on research objective, field, and dataset
- Common steps include standardising capitalisation, removing URLs and symbols, stopword removal, tokenisation, stemming, and lemmatization
- Stopwords include words like "a," "the," "of," "an" that don't add meaning to the dataset

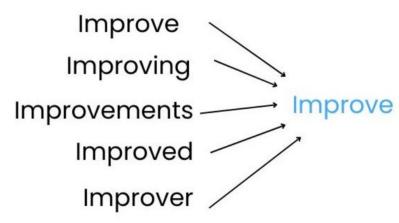






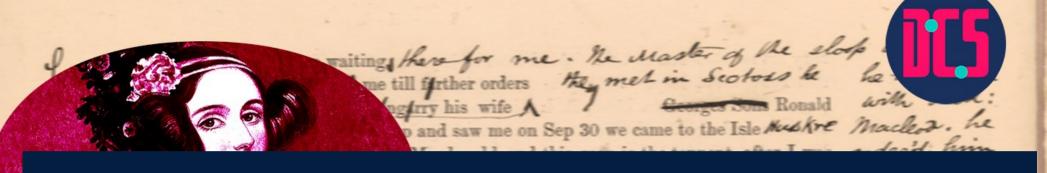
Stemming & Lemmatization





TURING





DEMO 6 8 7

time.

ciety sent a minister here I have given him a much that the sure I have given him the sur

0/t 0) 98



Finding Text Sources

- Libraries NLS Data Foundry (data.nls.uk)
- Project Gutenberg (gutenberg.org)
- Hathi Trust Digital Library (hathitrust.org)
- Websites Internet Archive (archive.org)'s
 Wayback Machine, UK Web
- archive (webarchive.org.uk)
- Newspaper archives (universities often subscribe to them!)





Research with NLTK

- Who is named in a text?
- What places are named in a text?
 - Chunking and Named Entity
 Recognition
- How does the vocabulary of an author change over time?
 - Lexical Diversity





Research with NLTK

- What are the common themes throughout a corpus?
 - Topic Modeling
- What attitudes are expressed in a corpus?
 - Sentiment Analysis
- What words occur near each other throughout a corpus? How does the meaning of a word change over time?
 - Word Embeddings







Next Week

- Research with NLTK on a corpus
 - NLTK with pandas (for tabular data)
 - NLTK with Altair (for data visualization)
- Regular Expression practice
- Cleaning messy text
- Resources for more text analysis practice

82 Hilda Jan: 20 173 It is agreat blessing and happiness to a natio when the Hing imployeth such a man as you are to chet and do for him who I'm personadid his the are and fear of God on Sum. job was a just man and aperfect and the the poor and oppressed and him that had none to he liem, affallerne for on in your office. I have the Monor to be your Relation and I know you have much interest with Lord Greange if you can make Peace for nie you know the promices that is to the Dear make of loveing my husband to much, he knowed very we that he was my job and now god his made him a vode lo Scourgeth me. * * * * * * * * *





Further Resources from CDCS

- Digital Method of the Month on Text Analysis
- Training Pathway for Text Analysis

82 Hilda Jan: 20 177 It is agreat blessing and happiness to a natio when the Hing imployeth such a man as you are to chet asul do for him who I'm personadid his the are and fear of God on Sum. job was a just man and aperfect and the the poor and oppressed and him that had none to he hiem, affallerne for on in your office. I have the Monor to be your Relation and I know you have much interest with Lord Greange if you can make Peace for of loveing my husband to much, he knowed very we that he was my job and now god his made him a vode lo Scourgeth me. * * * * * * * * *

