





#### **Course Topics**

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

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 and do for him who I'm personadid his the aw and feat of God on Sum. job was a just man and aperfect and the cause that he know not he searched out to deliver the poor and oppressed and him that had none to he liem, affallerne 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 jobol and now god his made him a vode lo Scourgeth me. \* \* \* \* \* \* \* \* \*





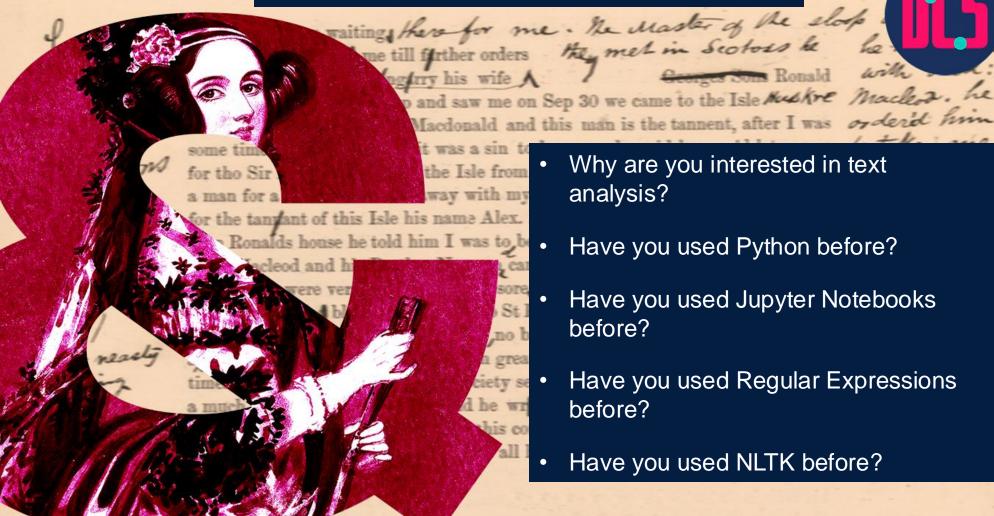
#### **Course Structure**

#### 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: 20 177 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 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. \* \* \* \* \* \* \* \* \*

## 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?





#### **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
questions

gr Kilda Jan: 20 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 him. job was a just man and aperfect 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 Beace make of loveing my husband to much, he knowed very we that he was my jobol and now god his made him a vode lo Scourgeth me. \* \* \* \* \* \* \* \* \*





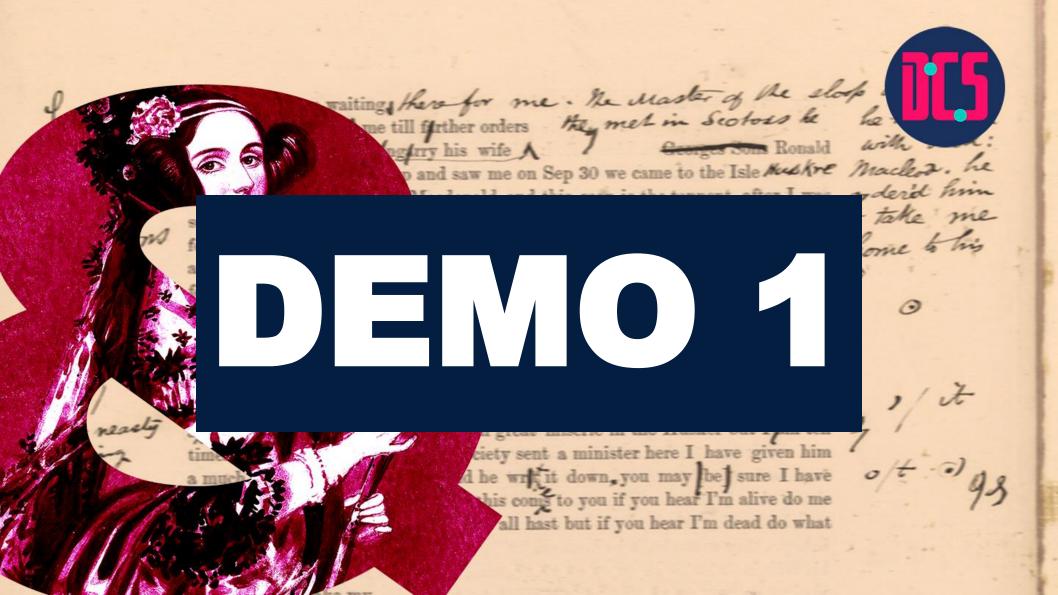
#### **Course Software**

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/introduction-and-installation





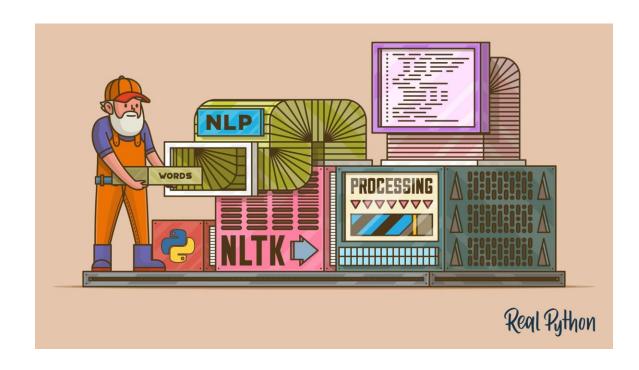




#### **NLTK**

Natural Language Toolkit

Natural language = human language
= "unstructured" data







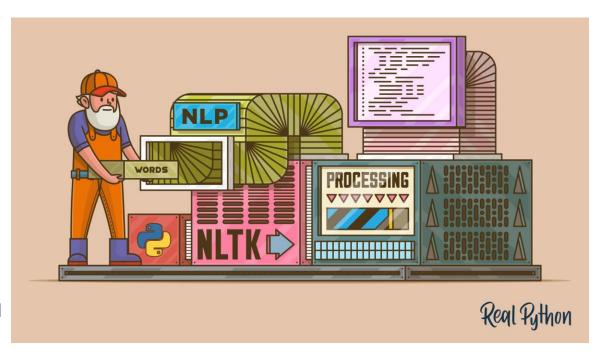


#### **NLTK**

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 Ofindous suito\*, Olivo\*fidelito\* à agilice « leuce funt in pas morieme bonous poges, ba Emite. 350 apandonin cade fin-de no observato bereires feignes (in mal., ved o e prire unidato neries. c Ex cogn. ex. q.o. 20 mits octorie e i collicité va ad bane cend venire posities qui nit de pagnament; obsobolo fine. cabente:qemutra pare nocere vobis

ei th auxilio bei fupare poteritis. maiores penas patiet of alif ad fillitu/ dine ille' q viu coburit. e Et cere rialfeger con miore pena punientur. velilli d erit ceterillad en pueri oc/ cifimado: vinent beo. f Et onnes auce dioce feti velecrati füt pepea illo rû t be pfectu frateli. Æa.XX

E vidi angela. Dicta banarioe b babylonie a antixpi a přeudo apřes iupponit bănationě tře biabelita pponit căm dre a tře d bás nabifre de illu poterit vincere falli d in buili aduciu potnit illu ligare cotta potnit, b Bhyff, Bhyff tenebola costa ipien vel ife biabol'de die fer nire pnittis trefrenat. i Bppben dirapeceaffe ofidit of fe imme a peo ille viabol'occidit. k Riganit en. Beiedi quill'i abrea valüs fidelibus issiedin getter aveza e zus fiedeltus vadorfinut igarisi in dies pätid-fied in ilis liganit fies futuri ppi, in ilis ipe eps advende liganit. I. Et misi et epe etce fa fiedelto i matis cepi voia ravort. Diegli in enangelisce pintra nit in porces. Is Signanit

Theo. delyra et buobemilib pedira 3 boftes quon nüer erat andeem miliü ablig ill' in ctaffe venerant antmollus pecdentes pftranerüt ex boftib maximam multi tudine a relidui conerfi füt ad fugam oc ciaffe criam illa magna pare a nrio fult restra oil à tobes: Et vidi vinum angein.i.priarcha bierofelymitanus. g- Stare in fole p eminetia vite bor erinerem ille Ecc. I. Bic fol refulges ficitierefulit in teplo pet. b Et cla.

e Dices ofb'ani. tc. que parella ad el'erbonanione facta aues fatts babu erat ad comedenda. o Et vidt be ftia.i.egyptifoldana, e Etreges terre.ibi inbiectos. f Et erercit co.co.in mari t in terra. g Ad faci endu plit ve a.ch rege balduino et epercitu fue. b Et appbefa e beitia. 6 ve.i.rex egypti v "ppbete falli qui in av fturia fina fuerūt coprebeli.na öriū el? go cogitamerat enenit. 1 Et & fecit

e f in lumine incelligiele feripeurup t fernose b f .i.meenenone. lum ftate in fole: e clamauit vo n für contiene i cele ett. b f. a.celetta con ce magna vices:oib amb g vo ct seplabant. a f p carpolica fide q coia c. b f d. beite operamont.

labat p mediű celi. Editer ogre a \$ -1.ad ipm ben d woben refectio erre. & ve ber lecremini i cruciare ipieg ra maiet de mineru. gamini ad cenă magnă vei: vt i maducetif carnes requi a carnel f undicit alimp. tribunop & carnes fortili & car/ nes cora a fedetia in ipis: a car

nes olin liberoză et fuop a pu fillozum et magnop. Et vidi be o ftia t reges terre t exercitul con u & "i.corra spin v fuos blafphomije v rebutari one v miraculia.

ogregatos ad faciedű pliű:cům s illo d ledebat i equo z cu ererci/ a f .i. commaliter punira.

befrie e d adorauerus imagine ! et". Willi miffi für bi ouo in fta. " f pprer cencheze. gnű ignis ardétis a fulpburis.

e et ceteri occili sut in gladio sede o tie sup equu qui peedit de ote f ipio. Et oce ques saturate sunt carnibus corum. La. XX

from. a s homane fem carne. u s pfie fcerero. b s'ecce porenia ieus tabiles. c s prierem fispra biabelli. tem De celo: Dabentes clá/ q qui Decalieto papa, a benrico impato

Mico. delvra

guine petorie et veferibit iftalans Gith ad ecelelia trifipba tom a quantu ad militante q veber quara por triupbati cons formari, fin gö ös in illa pfattöigne marris gandia reform tet ecelari ao aido viet eb ti förgo, ab alisa autem bingo-nia be (ado vietori, doss.) et Executia allelatis, pecela ri-iemphante, si ibdel polit. pr terri oliverit allelatis, omiliante.

ht Er cecidi ad pedes ci'te.p Boffe ditur o non folum bebet grane reddi Deo De fuis bificipe: fed etid fais mini ffrie p quoe reselat bolbus verirates vel alia bifficia fubministrat. † Et birit mibi vide ne fecenis ac.per B ar, guant boles fuperbi qui querunt ide bire ab aligs bonozari. at 100 bec vidi celum apertum 42.boc fraiter in telligi vider be ppo per es que fequun tur. Moraliter aur pot erponi be div bet er bono vicarto in fpuntibus cur inimodi funt prelati: vel tpalib" cuinfo modi funt principes terreni t con fabo din boni qui parati fiint ets obedire i licitis a boncitio lo fubditur: of Exercitus qui funt in celo fede

bantur eum in equis albis ac.i.in corpoibus mundis. by Et ibe calcus torentar ac.i. bonus rhi vicari punit iufte malefacta. † Et babet in ver ftimento a femoze fuo feriptum reg regum 12. qu find fe plures da regestes babet qui reges dici poffunt a regêdo de nar doc feripià in veftimento q in fe more fuo, femur eft intrinfech boirve, ftimenta ant extrinfech. In actibo aute intrinfecte talis fuptozis appet @ po-pula fibi fabdită vult bicto mo gaber \*\* f.i.emalier puter.

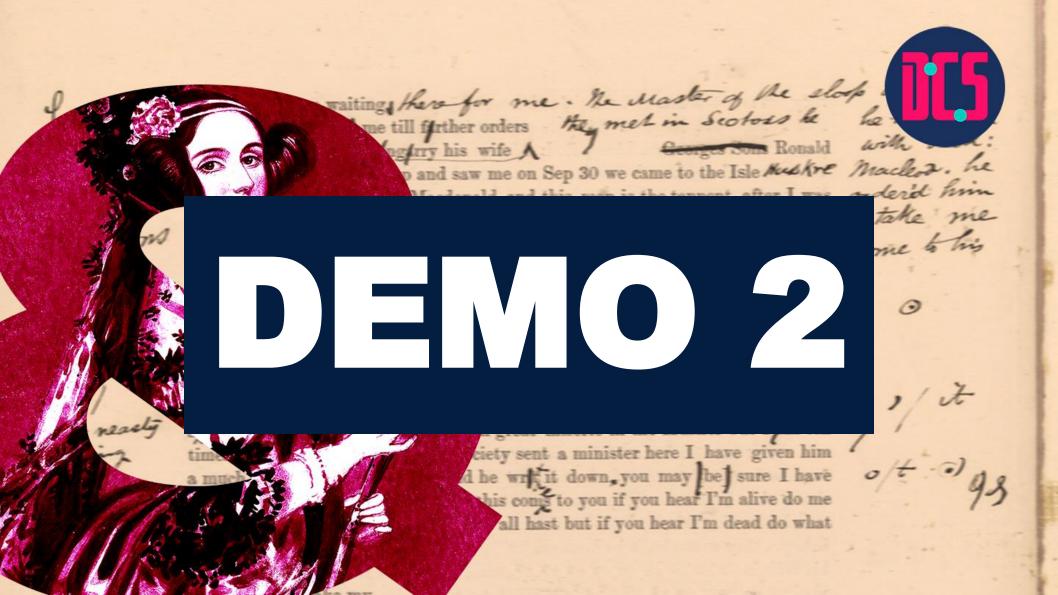
u cf. Et apprebela è befria: et but cf. Et vi. mà ange. flant i i coi illa pfeudo ppbeta: t qui fe i chi illa pfeudo proprieti characterem coi qui acceperuit characterem coi qui acceperuit characterem coi feu automi chi o pe automi chi coi buditi: c i De automi chi illa pfeudo ppbeta: t qui fe i chi illa pfeudo ppbeta: t qui fe i chi qui tuti i chi illa pfeudo ppbeta: t qui fe i chi illa pfeudo ppbeta: t qui feudo ppbeta: t qui f L'bemonibus à fant executores biuine inflicie in vanatis. o ant p aues fignificeri o emonoscipti p verba fatantori, in pebola feminatio. † Et madur cette carries rega et.i. obilitation in culpa io fequili: b† Ét apibenia é belia té.fequi: b† Zuffi tunt in that gnum ignis ardentis. Z vidi angelh. Defepta in ger e nerali a ípealí feptima plaga: 6

par veferibit ecelle pmotio no morto fedo ppianon iterara perfectitio ibi: Et vidi fedeo. Gelêdii in op pmam

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

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

www.cdcs.ed.ac.uk







#### 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.

(a Offichia amb. Offic fields) agiles e lettes finit in pa-motione bonop open. D Benite. Dis opandatin eade fi-de no observace benieve ficinetes file malt, ved ace pitto middle sucritis. Ex cognet. 4-q., Dullin footsis et isoliteit vit ad bane cend centre politicis qui nit bit pragnament, biabolo fice objectes a ministra et access and the constraints. cabente:qemutra pare nocere vobis

ei th auxilio bei fupare poteritis. maiores penas patiet of alijad filleus dine illie q viu coburit. e Et cere tili.fegers con miore pena punientur. veilil d grüt erteri.lad eb ouerfi oc effimado: vinent Deo. f Et omnes auce dioce feti velecrati füt pepea illo rû t be pfectu fratra. Ea.XX £a.XX

b babylonie a antrôl a pieudo aptor inpponit bănatione ipe biabelira pomit câm dre a tpe d băr nabifre de illu poterit vincere falli d in buili aductu potnit illa ligare città volutt giorificat cinde oino peftruere potnit, b Bbyff. Bbyff teneboofs coda ipicy vel ife biabol de año le-nic emitti trefrenat. i Bppben dira-peccasse ondir on se imane a peo ille viabol'occidit. k Riganit en. Beiedi quill'i abrea valüs fidelibus biabol fint ligar lic in ithis patib ifed in illis liganit (pes futuri rpi, in illis the the advented ligant. I Ermifit eti.ge eier etiche in cuangelio: qui intra ri atroct. b legit in cuangelio: qui intra nit in porceo. (3 Dignanit

Tico. delvra

et buobomilibopedira 5 boftes quon claffe venerant animolius peedentes pitranerat ex boftib maximam multi tudine a relidui conerfi füt ad fugam oc ciaffe criam illa magna pare a nrio fult restra oil à tobes: Et vidi vinum angein.i.priarcha bierofelymitanus. g- Stare in fole p eminetia vite bor erinerem ille Ecc. I. Bic fol refulges ficitierefulit in teplo pet. b Et cla.

e Dices ofb'ani. tc.qap ogreffi ad erat ad comedenda. o Et vidt be ftia.i.egyptifoldana, e Etreges terre.ibi inbiectos. f Et erercit ereitu fue. b Eraphblia beitig. ve.i.rex egypti v ppbete falfi qui in av fturia fina fuerūt coprebeit.na öriā el? go cogitanerat enenit. 1 Et d fecit

e f in lumine incelligiele feripeurup t fernose b f .i.meenenone. lum frate in fole: t clamauit vo e far coerfino l celo ett. b f. i. celeftia con seplabant, a f p carbolica fide q cois c. b f d. beite operamont.

labat p mediű celi. Editer ogre a \$ -1.ad ipm ben d woben refectio erre. & ve ber lecremini i cruciare ipiets ra maiets de minera. gamini ad cenă magnă vei: vt i maducetif carnes requi a carnel f undicit alimp. tribunop & carnes fortili & car/ nes cora a fedetia in ipis: a car

nes ofm liberozű et fuop a pu

fillozum et magnop. Et vidi be o ftia t reges terre t exercitul con a s.,i.cora spin v fues blafphonde v mbulari one v miraculis.

agregatos ad faciedú pliú:cum s illo d fedebati equo z cu exerci/ a f .i. commaliter punira.

cit signa coză spo quib" sedurit k & & si-qui in enm crediderane. cos qui acceperut characterem

befrie e d adorauerus imagine ! et". Chin miffi für bi ouo in fta/ " f pprer eencheze. gnú ignis ardêtis v fulpburis.

et ceteri occili fut in gladio fede . tie fup equi qui predit de ote f ipi" . Et oce ques faturate funt carnibus corum. La. XX

from a f hölliant fem carac.

†E vidi angelü Döfeende Pr.
a f přio ferro. b fecce potřná časí
tabios. c f prácem fipra otabelů.

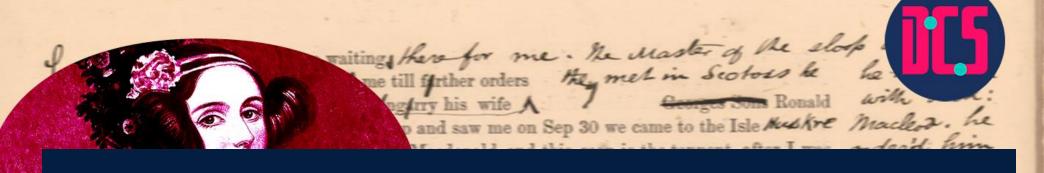
Mico. delvra guine petőzie et beferibit ifta lang Gefa ad ecelefia trifipba

tom a quantu ad militante q veber quara por triupbati cons formari, fin gö ös in illa pfattöigne marris gandia reform tet ecelari ao aido viet eb ti förgo, ab alisa autem bingo-nia be (ado vietori, doss.) et Executia allelatis, pecela ri-iemphante, si ibdel polit. pr terri oliverit allelatis, omiliante.

ht Er cecidi ad pedes ci'te.p Bofte ditur o non folum bebet grane reddi Deo De fuis bificipe: fed etid fais mini ffrie p quoe reselat bolbus verirates vel alia bifficia fubministrat. † Et birit mibi vide ne fecenis ac.per B ar, guant boles fuperbi qui querunt ide bite ab alije bonozari. # 120 bee vidi celum apertum se.boc fraiter in telligi vider be ppo per es que fequun tur. Moraliter aur pot erponi be div bet er bono vicarto in fpuntibus cur tufmodi funt prelati: vel tpalib" cutuf. modi funt principes terreni a con fubdin boni qui parati fiint ets obedire i licitis a boncitio lo fubditur: ot Et exercitus qui funt in celo fede

bantur eum in equis albis ac.i.in corpoibus mundis. by Et ibe calcus torentar ac.i. bonus rhi vicari punit iufte malefacta. † Et babet in ver ftimento a femoze fuo feriptum reg regum 12. qu find fe plures da regestes babet qui reges dici poffunt a regêdo de nar doc feripià in veftimento q in fe more fue.famur eft intrinfech boirves ftimenta aut errrinfech. In acrib aute intrinfecte talis fuptozis appet @ por pula fibi febdită vult victo mo gaber tu el'. Et apprebela è befria: et be nari. St. Et vi. en a ange. flamé in ci illa pieudo popoeta: t qui fe i clea gio poi intelligi daritas influi ci figna coza ipo quib' fedurit ki inclier eli ita admirabille. Bangel'igif ftans in fole elt pdicator conftanter p. dicas o obitinatos (eucritate inficie pinine.ideo firbdit: c† Be anibuf L'bemonibus à fant executores biuine infittic in panants. quit p aues fignificeri o emonoscipti p verba fatantori, in pebola feminatio. † Et madur cette carries rega et.i. obilitation in culpa io fequili: b† Ét apibenia é belia té.fequi: b† Zuffi tunt in that gnum ignis ardentis.

p E vidi angelü. Weiepta in ge-e nerali s ipēali ieptims plagatā pār deierībiš ecelie, pmotio no na s dindifigi pmo deierībiš dec pio motto.fco pounon iterara perfecutio ibi: Et vidi fedeo. Geledu th q pmam parrem a fedam in parte exponent all tem De celo: babentes cla/ q qui oc calinto papa. a benrico impato



## DEMO 3-5

d he writ it down, you may be sure I have given him to his come to you if you hear I'm alive do me all hast but if you hear I'm dead do what

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**

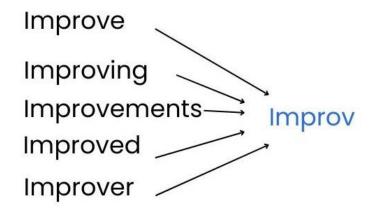
- Formatting text for analysis and removing extraneous informationWorkflows vary depending on research objective, field, and dataset
- Common steps include standardising capitalisation, removing URLs and symbols, stopword removal, tokenisation, stemming, and lemmatization
- Stopwords include words like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

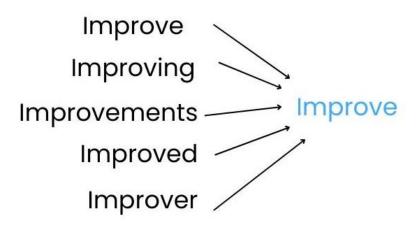






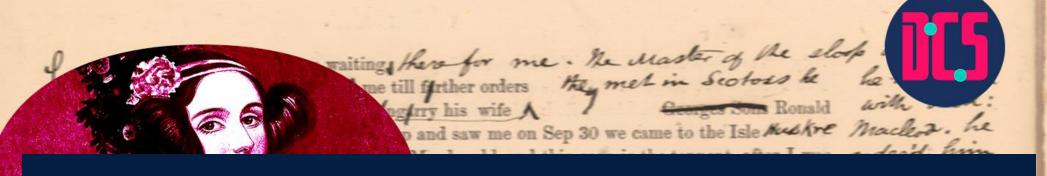
### **Stemming & Lemmatization**





TURING





# DEMO 6 8 7

time ciety sent a minister here I have given him a much a much a much his come to you if you hear I'm alive do me all hast but if you hear I'm dead do what

0/t 0) 93



#### **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

gr 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 and do for him who I'm personadid his the are and fear of God on him. 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 Greasure if you can make Peace for nie you know the promices that is to the Beace 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 and do for him who I'm personadid his the are and fear of God on him. job was a just man and aperfect and the cause that he know not be 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 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 jobol and now god his made him a vode lo Scourgeth me. \* \* \* \* \* \* \* \* \*

