Q1(a): Load the file in and use nltk.word_tokenizer() on it. Report the list of tokens that are produced from it and note any oddities that arise. Comment on these oddities and how they might be handled.

Oddities: 'it', '', 's' the nltk Python tokeniser split short form of words to many parts.

'https', ':', '//www.javatpoint.com/collections-in-java' it will be seen as two parts.

'5 \in ','1 \in ' both of them may be handled as a whole.

'23', '%', 'VAT' 'VAT' will be seen as a whole.

Q1(b): Now, take the output from normalization and run it through the pos-tagger.

```
postagged1=[('an', 'DT'), ('arraylist', 'NN'), ('preserves', 'VBZ'), ('the', 'DT'), ('order', 'NN'), ('of', 'IN'), ('elements', 'NNS'), (':', ':'), ('the', 'DT'), ('first', 'JJ'), ('element', 'NN'), ('of', 'DT'), ('clement', 'NN'), ('of', 'NN'), (
```

Q1(c): Report this output as your answer and highlight any inaccuracies that occur at this stage.

Inaccuracies: ('' ', 'VBZ'), ('s', 'RB'), ('o', 'UH'), ('|', 'NNP'), ('+', 'NNS'),('%', 'NN'), ('pos', 'JJ'), ('=', 'JJ')

Q2(a): Tokenize a new text-file (200 words) and the stem it using Porter Stemming. Report your answer and some of weird things that Porter Stemming does.

```
stemmed2= ['the', 'increas', 'use', 'of', 'robot', 'to', 'assist', 'surgeri', 'ha', 'enabl', 'contemporari', 'technolog', 'to', 'be', 'integr', 'into', 'medic', '.', 'although', 'robotic-assist', 'surgeri', 'wa', 'shown', 'to', 'be', 'competit', 'with', 'other', 'convent', 'surgic', 'procedur', 'and', 'appear', 'to', 'have', 'more', 'advantag', 'than', 'old', 'type', 'of', 'surgeri', ',', 'includ', 'less', 'blood', 'loss', ',', 'less', 'oper', 'trauma', ',', 'fewer', 'transfus', ',', 'less', 'risk', 'of', 'infect', ',', 'faster', 'recoveri', 'and', 'quicker', 'return', 'to', 'common', 'activ', ',', 'it', 'still', 'exist', 'critic', 'viewpoint', ',', 'in', 'which', 'the', 'advers', 'event', 'produc', 'by', 'surgic', 'robot', 'report', 'to', 'the', 'publicli', 'avail', 'maud', 'databas', '(', 'maintain', 'by' 'f.d.a', ')', 'from', '2000', 'to', '2013', 'is', 'perform', 'by', 'alemzadeh', 'et', 'al', '.', 'As', 'fig', '5', 'show', ',', 'the', 'overal', 'trend', 'of', 'injuri', 'and', 'death', 'event', 'in', 'the', 'annual', 'number', 'of', 'report', 'and', 'the', 'estim', 'rate', 'of', 'occurr', 'per', '100,000', 'oper', 'have', 'been', 'increas', 'from', '2004', 'to', '2013', '(', '2016', ')', '.', 'the', 'increas', 'report', 'of', 'event', 'may', 'occur', 'due', 'to', 'variou', 'reason', '.', 'what', 'we', 'could', 'discov', 'from', 'the', 'access', 'inform', 'is', 'that', 'the', 'greater', 'complex', 'of', 'the', 'process', ',', 'less', 'frequent', 'use', 'of', 'robot', 'machin', ',', 'and', 'lack', 'of', 'suffici', 'robot', 'expertis', 'in', 'these', 'area', '.', 'under', 'these', 'circumst', ',', 'there', 'are', 'two', 'perspect', 'that', 'we', 'can', 'explor', 'to', 'find', 'a', 'solut', 'to', 'solv', 'these', 'problem', '.']
```

Weird things: has-ha, was-wa, MAUDE-maud, database-databas, machine-machin, estimated-estim ...

Q2(b): Tokenize the new text-file and then lemmatize it using WordNet Lemmatizer; note you may have to pos-tag the sentences first and then convert the tags to make this work. Report the result of these steps and point out some of the things that look wrong.

```
postagged2= [('The', 'DT'), ('increased', 'VBN'), ('use', 'NN'), ('of', 'IN'), ('robots', 'NNS'), ('to', 'TO'), ('assist', 'VB'), ('surgery', 'NN'), ('has', 'VBZ'), ('conabled', 'VBN'), ('contemporary', 'JJ'), ('technology', 'NN'), ('to', 'TO'), ('be', 'VB'), ('integrated', 'VBN'), ('into', 'IN'), ('medication', 'NN'), ('to', 'TO'), ('be', 'VB'), ('competitive', 'JJ'), ('intity', 'IN'), ('ofter', 'JJ'), ('conventional', 'JJ'), ('yrocedures', 'NBS'), ('and', 'CC'), ('appeared', 'VBD'), ('to', 'TO'), ('have', 'VB'), ('mone', 'JJR'), ('lood', 'NN'), ('advanse', 'NNS'), ('than', 'IN'), ('of', 'JN'), ('of', 'IN'), ('of', 'NN'), (',',','), ('integrated', 'VBD'), ('to', 'TO'), ('have', 'VB'), ('mone', 'JJR'), ('to', 'TO'), ('appeared', 'VBD'), ('to', 'TO'), ('have', 'VB'), ('mone', 'JJR'), ('to', 'TO'), ('advanse', 'NNS'), ('than', 'IN'), ('of', 'IN'), ('ins', 'NN'), (',',','), ('less', 'JJR'), ('transfusions', 'NNS'), (',',','), ('less', 'JJR'), ('trist', 'NN'), ('n', 'NN
```

lemmatized2= ['The', 'increase', 'use', 'of', 'robot', 'to', 'assist', 'surgery', 'have', 'enable', 'contemporary', 'technology',
'to', 'be', 'integrate', 'into', 'medication', '.', 'Although', 'Robotic-assisted', 'surgery', 'be', 'show', 'to', 'be',
'competitive', 'with', 'other', 'conventional', 'surgical', 'procedure', 'and', 'appear', 'to', 'have', 'more', 'advantage',
'than', 'old', 'type', 'of', 'surgery', ',' include', 'less', 'blood', 'loss', ',' less', 'operative', 'trauma', ',', 'few',
'transfusion', ',', 'less', 'risk', 'of', 'infection', ',', 'fast', 'recovery', 'and', 'quick', 'return', 'to', 'common',
'activity', ',' it', 'still', 'exist', 'critical', 'viewpoint', ',', 'in', 'which', 'the', 'adverse', 'event', 'in', 'robotic',
'surgery', 'be', 'the', 'main', 'problem', 'A', 'comprehensive', 'analysis', 'of', 'adverse', 'event', 'produce', 'by',
'surgical', 'robot', 'report', 'to', 'the', 'publicly', 'available', 'MAUDE', 'database', '(', 'maintain', 'by', 'F.D.A', ')',
'from', '2000', 'to', '2013', 'be', 'perform', 'by', 'Alemzadeh', 'et', 'al', '.', 'As', 'fig', '5', 'show', ',', 'the',
'overall', 'trend', 'of', 'injury', 'and', 'death', 'event', 'in', 'the', 'annual', 'number', 'of', 'report', 'and', 'the',
'estimate', 'rate', 'of', 'occurrence', 'per', '100,0000', 'operation', 'have', 'be', 'increase', 'from', '2013',
'(', '2016', ')', '.', 'The', 'increased', 'reporting', 'of', 'event', 'may', 'occur', 'due', 'to', 'various', 'reason', '.',
'What', 'we', 'could', 'discover', 'from', 'the', 'accessible', 'information', 'be', 'that', 'the', 'great', 'complexity', 'of',
'the', 'process', ',' 'less', 'frequent', 'use', 'of', 'robotic', 'machine', ',', 'and', 'lack', 'of', 'sufficient', 'robotic',
'expertise', 'in', 'these', 'area', '.', 'Under', 'these', 'circumstance', ',', 'there', 'be', 'two', 'perspective', 'that', 'we',
'can', 'explore', 'to', 'find', 'a', 'solution', 'to', 'solve', 'these', 'problem', '.']

Q2(c): Compare the outputs from Porter Stemming and the Lemmatisation of the same file. Which do you think is the best to use and why?

After comparing the outputs from these two methods. I think lemmatisation is the best to use. Because lemmatisation is the process of converting a word to its base form while Stemming is just the removal of morphological affixes and it works off mixture of specific rules. For instance:

"database" was converted to databas in porter stemming. But it remained in lemmatisation. lemmatization considers the context and converts the word to its meaningful base form, whereas stemming just removes the last few characters, which may lead to misunderstanding and spelling errors.

Q3: Finally, choose a remote webpage and extract key text content from it. Install the packages you need and then parse it using BeautifulSoup. Try to get to a point where you can extract one of its XML/HTML parts (e.g., title, summary, body)

title: <title>School of Computer Science</title>

body: Skip to main contentSchool of Computer ScienceToggle navigationOur SchoolAbout UsContactTestimonialsUseful LinksStudy With UsUndergraduatePostgraduateResearch DegreesInternational StudentsResearch CentresResearch InterestsResearch ThemesPublicationsPeopleAcademic StaffSupport StaffPhD StudentsQuick LinksWebsite Login (Staff Only)CS MoodLeUCD BrightspaceCS GradingSchool Safety StatementCS Plagiarism PolicySupport Centre (CSSC) Free tutoringFourth Year ProjectsTech SupportCS GitLabStudent Support OfficeRoom BookingsUCD Academic RegulationsAthena SWANEU SWIM ProjectSearch formSearchUndergraduatePostgraduateInternational StudentsResearchNews and EventsArtificial Intelligence Innovation Hub CeADAR Secures €12 million in Funding from Enterprise IrelandUpcoming seminar: Urbashi MitraUCDCS Best Paper ICCBR ConferenceShow moreBlogSecurity, Privacy and Digital Forensics in the Cloud2019-08-30A new book edited by UCD's Dr Nhien-An Le-Khac has been listed the Best New Digital Forensics Book To Read In 2019 according to...Chidubem Iddianozie: PhD student and GitHub Ambassador in UCD2019-09-03Chidubem is a second year PhD Student in the UCD School of Computer Science, working on semantic inference for geo-spatial data. He has been a GitHub Campus Expert since February 2019....UCD CS PhD student selected to attend the Heidelberg Laureate Forum. This is the second time Zahid has been selected to attend this annual event...New Project: Evidence-Based Decision Support for Real-Estate Investment2019-05-29With global real estate transactions predicted to surpass \$1 trillion per year in 2020 how can investors assess their return and mitigate risk?...UCD projects celebrate Europe Day.019-05-10UCD Computer Science staff and faculty in two ERDF-funded projects joined their colleagues around the university today to celebrate Europe Day. The two projects focus on coastal water quality, with particular reference to recreational use of our coasts. The projects, SMIM and ACCLIMATIZE, are fun...Critical thinking and data ethics in UCD