

Assignment Project Exam Help

Text Pre-Processing — 1

<https://powcoder.com>

Faculty of Information Technology, Monash University, Australia

Add WeChat powcoder

FIT5196 week 4

Assignment Project Exam Help

1 Basic Tasks in Text Preprocessing

- Tokenization
- Case Normalization
- Stopping — Remove Stop Words
- Stemming & Lemmatisation
- Sentence Segmentation

<https://powcoder.com>

Add WeChat powcoder

Text is everywhere!

- A large amount of text data available in different forms. For example,

- ▶ e-books & e-magazines
- ▶ online newspapers
- ▶ blogs & tweets
- ▶ online product reviews
- ▶ emails
- ▶ Medical reports and articles
- ▶ Research papers published by different conferences

- Various text resources

- ▶ Online data repositories: UCI machine learning repository, Linguistic Data Consortium
- ▶ NLTK: the built-in datasets
- ▶ Web: Crawl text data by yourself!

- How to use automatic approaches to analyse the text syntactically and semantically?

- The goal of text analysis: provide understanding of how the text is processed without having a human read it.
- The capability of computers

- ▶ What a computer can do:

- Examine the individual characters in each word, and how those words are arranged.

- ▶ What a computer cannot do:

- Know what the information is communicated by the text syntactically and semantically.

- Syntax v.s. Semantics

- ▶ Syntax: the structure of language, e.g., grammar rules

- How individual words are composed to make well-formed sentences and paragraphs.

- ▶ Semantics: the meaning of the individual words within the surrounding context

- Understand the theme of a given text fragment.

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Examples of Text Analysis Tasks

- Analyse sentence structure — e.g., syntactic parsing and dependency

parsing

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

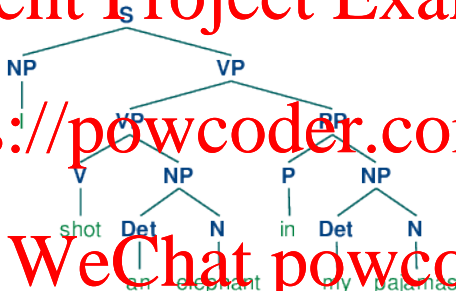


Figure: Figure from the NLTK book

Examples of Text Analysis Tasks

- Topic modelling

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
TO	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$125 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

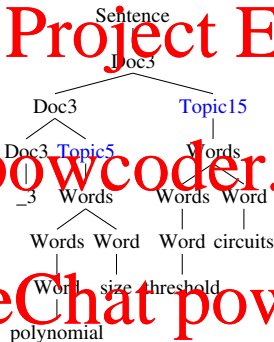
Examples of Text Analysis Tasks

- Learn topical phrases? — Topical collocation

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



- Is "white house" a topical collocation?

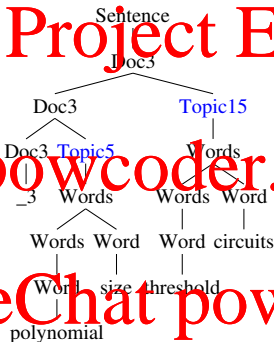
Examples of Text Analysis Tasks

- Learn topical phrases? — Topical collocation

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



- Is "white house" a topical collocation?
 - In a real-estate context: compositional phrase
 - In a political context: topical collocation

Examples of Text Analysis Tasks

- Break down document into topically coherent chunks — Text segmentation

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Figure: A 21-paragraph science news article, called Stargazers, from Hearst, 1997. The main topic is the existence of life on earth and other planets.

- 1-3 *Intro - the search for life in space*
- 4-5 *The moon's chemical composition*
- 6-8 *How early earth-moon proximity shaped the moon*
- 9-12 *How the moon helped life evolve on earth*
- 13 *Improbability of the earth-moon system*
- 14-16 *Binary/trinary star systems make life unlikely*
- 17-18 *The low probability of non-binary/trinary systems*
- 19-20 *Properties of earth's sun that facilitate life*
- 21 *Summary*

- Text data always appears in an unstructured form.

igment Project Exam He

Text Data in a Structured Form

- Goal: manipulate and convert the free language text into structured form.

Assignment Project Exam Help

https://

Add What new coder

```
0,the
1,newlyreleased
2,katana
3,lx
4,is
5,successor
6,to
7,ii
8,and
9,latest
10,addition
11,sanyos
12,line
13,it
14,features
15,a
16,new
17,narrower
18,design
19,with
20,amirrored
21,finish
22,hidden
23,oled
24,outer
25,display
26,though
```

```
0:0,2,3,6,2,2,15:2,8:1,74,2,1,5:1,14,1,5,1,1,2,1,15,1,1,168,1,193,4,52,2,4,3,90:1,21,335:1,337:1,387,1,450,1,5,1,1,32:1,
1,1041:2,1215:1,1216:1,1217:2,1218:1,1219,1,1220:1,1221:1,1222:1,1223:1,1224:1,1225:1,1226:1,1227:1,1228:1,1229:1,1230:1,1231:1,1232:
1,1233:1,1234:1
0:2,4:1,8:2,36:1,50:1,146:1,188:1,239:1,410:1,928:1,1227:1,1235:1,1236:1,1237:1,1238:1,1239:1
3:1,805:1,1240:1,1241:1,1242:1,1243:1,1244:1,1245:1,1246:1,1247:1,1248:1,1249:1,1250:1,1251:1,1252:1,1253:1,1254:1
0:6,6:2,8:2,13:1,15:2,16:1,41:1,71:2,106:3,338:1,543:1,558:1,714:1,741:1,824:1,1051:1,1162:1,1255:1,1256:1,1257:1,1258:1,1259:1,1260
:1,1261:1,1262:1,1263:1,1264:1,1265:1,1266:1,1267:1,1268:1,1269:1,1270:1,1271:1,1272:1
0:10,4:2,6:2,8:3,13:3,15:4,19:3,25:1,31:2,36:6,45:1,46:1,56:2,74:3,84:1,91:1,95:1,106:6,114:1,142:1,146:1,159:1,164:2,177:1,180:1,25
0:1,252:2,2,9:1,281:1,284:1,291:1,358:4,362:1,437:1,532,1,558:1,571:1,574:1,625:1,646:1,649:1,680:1,945:1,967:1,1046:1,1095:1,1096:1
1,1244:1,1237:1,1238:1,1239:1,1240:1,1241:1,1242:1,1243:1,1244:1,1245:1,1246:1,1247:1,1248:1,1249:1,1250:1,1251:1,1252:1,1253:1,1254:1
1,1255:1,1256:1,1257:1,1258:1,1259:1,1260:1,1261:1,1262:1,1263:1,1264:1,1265:1,1266:1,1267:1,1268:1,1269:1,1270:1,1271:1,1272:1
1,1273:1,1274:1,1275:1,1276:1,1277:1,1278:1,1279:1,1280:1,1281:1,1282:1,1283:1,1284:1,1285:1,1286:1,1287:1,1288:1,1289:1
1,1290:1,1291:1,1292:1,1293:1,1294:1,1295:2,296,297,298:1,299:1,1,90:1,101:1,112:1,13:3,1,1304:1,1305:1,1306:1,1307:1,1308
:1,1309:1,1310:1,1311:1,1312:1,1313:1,1314:1,1315:1,1316:1,1317:1,1318:1,1319:1,1320:1,1321:1,1322:1,1323:1,1324:1,1325:1,1326:1
0:1,6:1,42:1,1,1,129:1,130:2,131:1,132:1,133:1,134:1,135:1,136:1,137:1,571:1,1300:1
0:15,4:1,6:5,8:5,15:3,19:1,25:2,36:1,46:1,51:1,82:1,87:1,106:1,114:1,120:3,122:1,123:1,164:1,171:1,176:2,178:1,196:1,222:1,238:1,239
2,254:2,255:1,261:2,269:1,284:2,297:1,342:1,369:2,381:1,386:2,442:2,473:1,608:1,622:1,654:1,686:1,762:1,779:1,796:2,945:1,1101:1,11
33:2,1179:2,1182:1,1300:1,1301:1,1303:1,1306:1,1307:1,1309:3,1327:1,1328:1,1329:1,1330:1,1331:1,1332:1,1333:1,1334:1,1335:1,1336:1,1
337:1,1338:1,1339:1,1340:1,1341:1,1342:1,1343:1,1344:1,1345:1,1346:1,1347:1,1348:1,1349:1
0:8,6:1,8:2,15:1,36:3,37:1,41:1,84:1,95:1,106:3,215:1,248:1,290:2,297:1,358:3,402:1,558:1,663:1,686:1,693:1,742:1,759:1,801:1,806:1,
8,1,1,124:1,1101:1,1290:1,1300:1,1309:4,1350:1,1351:1,1352:1,1353:1,1354:1,1355:1,1356:1,1357,1,1358:1,1359:1,1360:1,1361:1,1362:1,
13,1,1,1,664:1,1365:1,1,661:1
0:1,4:4,8:5,15:3,19:1,25:2,36:1,46:1,48,1,56,3,84:1,10:3,118:1,1,138:1,1,75:1,1,6:1,19:2,206:1,207,1,208:1,211:1,215:1,218:1,234:1,239:2
2,249,2,251:1,261,1,269,1,284:1,297,1,318:1,44:1,518,1,55,1,571,1,5,1,135,1,930:1,129,1,136:1,1307:1,1309:1,1367:2,1368:1,1369:1
1,1370:1,1371:1,1372:1,1373:1,1374:1,1375:1,1376:1,1,1377:1,1378:1,1379:1,1380:1,1381:1,1382:1,1383:1,1384:1
0:6,4:3,6:4,8:5,13:2,15:5,19:2,36:1,37:1,45:1,51:2,50:2,82:1,84:1,105:1,106:1,111:3,114:1,122:1,126:1,131:1,142:2,159:2,175:3,176:1,
178:1,196:1,197:1,202:1,215:1,218:1,222:2,226:1,246:1,248:2,252:1,290:1,358:4,381:1,403:2,436:1,518:1,574:1,663:2,717:1,743:1,789:1,
797:1,801:1,887:1,921:1,950:1,1003:1,1014:2,1239:1,1298:1,1309:6,1343:1,1344:1,1363:1,1367:1,1369:1,1385:2,1386:2,1387:1,1388:1,1389
1,1390:1,1391:1,1392:1,1393:1,1394:1,1395:2,1396:1,1397:1,1398:1,1399:1,1400:1,1401:1,1402:1,1403:1,1404:1,1405:1,1406:1,1407:1,140
8:1,1409:1,1410:1,1411:1,1412:1,1413:1
0:3,8:1,15:1,36:2,51:1,84:1,111:2,114:1,122:2,226:1,252:1,264:1,310:1,358:1,505:1,521:1,574:1,663:1,887:1,946:1,1264:1,1300:1,1307:1
1,1309:1,1414:1,1415:1,1416:1,1417:1,1418:1,1419:1,1420:1,1421:1,1422:1,1423:1,1424:1,1425:1,1426:1
0:2,4:2,6:3,8:3,13:2,15:1,51:1,52:1,73:2,106:2,111:1,122:2,142:1,146:1,173:1,218:1,252:1,335:1,358:4,457:1,580:1,654:1,670:1,789:1,7
96:1,1181:1,1182:1,1308:1,1309:3,1375:1,1386:3,1423:2,1427:2,1428:1,1429:2,1430:1,1431:1,1432:1,1433:2,1434:1,1435:1,1436:1,1437:1,1
438:1,1439:1,1440:1,1441:1,1442:1,1443:2,1444:1,1445:1
@
@
```

Basic Tasks in Text Preprocessing

- Tokenisation
- Case Normalisation:
- Stopping
- Stemming & Lemmatisation
- Sentence segmentation

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Basic Tasks in Text Preprocessing —Tokenisation

- Tokenisation: the process of breaking a stream of text into tokens.
 - ▶ Text is usually represented as sequences of characters by computers.

"A data wrangler is the person performing the wrangling tasks."

- ▶ Most natural language processing (NLP) and text mining algorithms can only operate on tokens.

["A", "data", "wrangler", "is", "the", "person", "performing", "the", "wrangling", "tasks"]

- Challenging issues:

- ▶ Periods in Abbreviations
 - Common acronyms with periods: U.K., U.N. etc.
 - Other abbreviations with a similar pattern: P.M., A.M., i.e., etc.
- ▶ Currency and Percentages
 - Different currencies: \$10,000.00, £10,000.00/00, AUD100, EUR10.555 and CNY555.55.
 - Percentages: 23%, 23.23% and 100.00%
- ▶ Hyphens and Apostrophes
 - Hyphens: "co-operate", "co-education" and "pre-process"
 - Apostrophes: "don't", "she'll"

Looking at the Jupyter Notebook!

Basic Tasks in Text Preprocessing — Case Normalisation

- Capitalisation helps readers differentiate, for example, between nouns and proper nouns

- ▶ Common nouns: writer, teacher, cookies, . . .

- ▶ Proper noun: Herman Melville, Snoopy, University of Melbourne, . . .

- Case normalisation: covert all the words into either uppercase or lowercase words

- ▶ "data" v.s. "Data"

- Case normalisation is not always needed.

- ▶ Information Retrieval: "Data Wrangler" v.s. "data wrangler"

- ▶ Named Entity Recognition: One would better keep capitalised words left as capitalised.

- One function to finish case normalisation: `lower()`

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



Basic Tasks in Text Preprocessing — Removing Stop Words

- Stop words — words that are extremely common and carry little lexical content. They are often function words in English. For example,

- ▶ articles (e.g., "a", "the", and "an"),
- ▶ pronouns (e.g., "he", "him", and "they"),
- ▶ particles (e.g., "well", "however" and "thus")

- Stop words usually refer to the most common words in a language. The general strategy for determining whether a word is a stop word or not is to compute its total number of appearances in a corpus.

- Why should we remove stop words?

- ▶ Stop words usually appear to be of little value and have little impact on the final results, as the presence of stop words in a text document does not really help distinguishing it from other documents.
- ▶ Failing to remove those common words could lead to skewed analysis results. For example,
 - Email analysis: remove headers (e.g., "Subject", "To", and "From"), remove a lengthy legal disclaimer, ...



Stemming & Lemmatisation

- Should we keep word forms like "educate", "educated", "educating", and "educates" separate or to collapse them?

- Stemming: the process of identification and removal of prefixes, suffixes, and pluralisation, which leaves you with a stem.

- ▶ 'watches -> watch'
- ▶ 'parties -> party'
- ▶ 'carrying -> carry'
- ▶ 'loving -> lov'

- Lemmatization: a more advanced form of stemming that makes use of, for example, the context surrounding the words, an existing vocabulary, morphological analysis of words and other grammatical information (e.g., part-of-speech tags) to determine the basic or dictionary form of a word, which is known as the lemma. Use `from nltk.stem import WordNetLemmatizer`

- ▶ "meeting" + (POS = 'v') → "meet"
- ▶ "meeting" + (POS = 'n') → "meeting"

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder

Basic Tasks in Text Preprocessing — Sentence Segmentation







- Sentence segmentation — a challenging problem in natural language processing, which is about deciding where sentences begin and end.
- Challenge: punctuation marks are often ambiguous
 - ▶ Is something ending with one of the following punctuations “.”, “!”, “?” ?
 - ▶ Does a period always indicate sentence boundaries?
 - Some periods occur as part of abbreviations, monetary numerals, percentages, decimal point, or an email address.
- The NLTK's Punkt Sentence Tokenizer was designed to split text into sentences “by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences.”
- Any other cues that can be used to identify a sentence boundary?

Summary: what do you need to do in this week?

- Download and read the materials provided in Moodle, and also read the recommended reading materials associated with each chapter.

Assignment Project Exam Help

4. Reference Reading Materials

1. "[Tokenization](#)" .
2. "[Processing Raw Text](#)", chapter 3 of of "Natural Language Processing with Python".
3. "[Tokenization](#)" An IBM blog on tokenization. It gives a detailed discussion about word tokenization and its challenges .
4. "[Sampling and Lemmatization](#)" .
5. "[Dropping common terms: stop words](#)" .
6. "[Corpus-Based Work](#)", Chapter 4 of "Foundations of statistical natural language processing" by Christopher D. Manning .
7. "[Testing out the NLTK sentence tokenizer](#)"
8. "[Accessing Text Corpora and Lexical Resources](#): Chapter 2 of "Natural Language Processing with Python" By Steven Bird, Ewan Klein & Edward Loper .
9. "[Corpus Readers](#)": An NLTK tutorial on accessing the contents of a diverse set of corpora.

https://powcoder.com

Add WeChat powcoder