dtm_example

July 2, 2018

1 DTM Example

In this example we will present a sample usage of the DTM wrapper. Prior to using this you need to compile the DTM code yourself or use one of the binaries.

This tutorial is on Windows. Running it on Linux and OSX is the same.

In this example we will use a small already processed corpus. To see how to get a dataset to this stage please take a look at Gensim Tutorials

```
In [1]: import logging
        import os
        from gensim import corpora, utils
        from gensim.models.wrappers.dtmmodel import DtmModel
        import numpy as np

        if not os.environ.get('DTM_PATH', None):
            raise ValueError("SKIP: You need to set the DTM path")
```

First we wil setup logging

```
In [2]: logger = logging.getLogger()
        logger.setLevel(logging.DEBUG)
        logging.debug("test")
```

Now lets load a set of documents

```
In [3]: documents = [[u'senior', u'studios', u'studios', u'studios', u'creators', u'award',
        u'mobile', u'currently', u'challenges', u'senior', u'summary', u'senior', u'motivated',
        u'creative', u'senior', u'performs', u'engineering', u'tasks', u'infrastructure',
        u'focusing', u'primarily', u'programming', u'interaction', u'designers', u'engineers',
        u'leadership', u'teams', u'teams', u'crews', u'responsibilities', u'engineering',
        u'quality', u'functional', u'functional', u'teams', u'organizing', u'prioritizing',
        u'technical', u'decisions', u'engineering', u'participates', u'participates',
        u'reviews', u'participates', u'hiring', u'conducting', u'interviews', u'feedback',
        u'departments', u'define', u'focusing', u'engineering', u'teams', u'crews',
        u'facilitate', u'engineering', u'departments', u'deadlines', u'milestones',
        u'typically', u'spends', u'designing', u'developing', u'updating', u'bugs',
        u'mentoring', u'engineers', u'define', u'schedules', u'milestones', u'participating',
        u'reviews', u'interviews', u'sized', u'teams', u'interacts', u'disciplines',
        u'knowledge', u'skills', u'knowledge', u'knowledge', u'xcode', u'scripting',
        u'debugging', u'skills', u'skills', u'knowledge', u'disciplines', u'animation',
        u'networking', u'expertise', u'competencies', u'oral', u'skills', u'management',
        u'skills', u'proven', u'effectively', u'teams', u'deadline', u'environment',
        u'bachelor', u'minimum', u'shipped', u'leadership', u'teams', u'location', u'resumes',
        u'jobs', u'candidates', u'openings', u'jobs'], [u'maryland', u'client', u'producers',
        u'electricity', u'operates', u'storage', u'utility', u'retail', u'customers',
        u'engineering', u'consultant', u'maryland', u'summary', u'technical', u'technology',
        u'departments', u'expertise', u'maximizing', u'output', u'reduces', u'operating',
        u'participates', u'areas', u'engineering', u'conducts', u'testing', u'solve',
```

u'supports', u'environmental', u'understands', u'objectives', u'operates',
u'responsibilities', u'handles', u'complex', u'engineering', u'aspects', u'monitors',
u'quality', u'proficiency', u'optimization', u'recommendations', u'supports',
u'personnel', u'troubleshooting', u'commissioning', u'startup', u'shutdown',
u'supports', u'procedure', u'operating', u'units', u'develops', u'simulations',
u'troubleshooting', u'tests', u'enhancing', u'solving', u'develops', u'estimates',
u'schedules', u'scopes', u'understands', u'technical', u'management', u'utilize',
u'routine', u'conducts', u'hazards', u'utilizing', u'hazard', u'operability',
u'methodologies', u'participates', u'startup', u'reviews', u'pssr', u'participate',
u'teams', u'participate', u'regulatory', u'audits', u'define', u'scopes', u'budgets',
u'schedules', u'technical', u'management', u'environmental', u'awareness',
u'interfacing', u'personnel', u'interacts', u'regulatory', u'departments', u'input',
u'objectives', u'identifying', u'introducing', u'concepts', u'solutions', u'peers',
u'customers', u'coworkers', u'knowledge', u'skills', u'engineering', u'quality',
u'engineering', u'commissioning', u'startup', u'knowledge', u'simulators',
u'technologies', u'knowledge', u'engineering', u'techniques', u'disciplines',
u'leadership', u'skills', u'proven', u'engineers', u'oral', u'skills', u'technical',
u'skills', u'analytically', u'solve', u'complex', u'interpret', u'proficiency',
u'simulation', u'knowledge', u'applications', u'manipulate', u'applications',
u'engineering', u'calculations', u'programs', u'matlab', u'excel', u'independently',
u'environment', u'proven', u'skills', u'effectively', u'multiple', u'tasks',
u'planning', u'organizational', u'management', u'skills', u'rigzone', u'jobs',
u'developer', u'exceptional', u'strategies', u'junction', u'exceptional', u'strategies',
u'solutions', u'solutions', u'biggest', u'insurers', u'operates', u'investment'],
[u'vegas', u'tasks', u'electrical', u'contracting', u'expertise', u'virtually',
u'electrical', u'developments', u'institutional', u'utilities', u'technical',
u'experts', u'relationships', u'credibility', u'contractors', u'utility', u'customers',
u'customer', u'relationships', u'consistently', u'innovations', u'profile',
u'construct', u'envision', u'dynamic', u'complex', u'electrical', u'management',
u'grad', u'internship', u'electrical', u'engineering', u'infrastructures', u'engineers',
u'documented', u'management', u'engineering', u'quality', u'engineering', u'electrical',
u'engineers', u'complex', u'distribution', u'grounding', u'estimation', u'testing',
u'procedures', u'voltage', u'engineering', u'troubleshooting', u'installation',
u'documentation', u'bsee', u'certification', u'electrical', u'voltage', u'cabling',
u'electrical', u'engineering', u'candidates', u'electrical', u'internships', u'oral',
u'skills', u'organizational', u'prioritization', u'skills', u'skills', u'excel',
u'cadd', u'calculation', u'autocad', u'mathcad', u'skills', u'skills', u'customer',
u'relationships', u'solving', u'ethic', u'motivation', u'tasks', u'budget',
u'affirmative', u'diversity', u'workforce', u'gender', u'orientation', u'disability',
u'disabled', u'veteran', u'vietnam', u'veteran', u'qualifying', u'veteran', u'diverse',
u'candidates', u'respond', u'developing', u'workplace', u'reflects', u'diversity',
u'communities', u'reviews', u'electrical', u'contracting', u'southwest', u'electrical',
u'contractors'], [u'intern', u'electrical', u'engineering', u'idexx', u'laboratories',
u'validating', u'idexx', u'integrated', u'hardware', u'entails', u'planning', u'debug',
u'validation', u'engineers', u'validation', u'methodologies', u'healthcare',
u'platforms', u'brightest', u'solve', u'challenges', u'innovation', u'technology',
u'idexx', u'intern', u'idexx', u'interns', u'supplement', u'interns', u'teams',
u'roles', u'competitive', u'interns', u'idexx', u'interns', u'participate',
u'internships', u'mentors', u'seminars', u'topics', u'leadership', u'workshops',
u'relevant', u'planning', u'topics', u'intern', u'presentations', u'mixers',
u'applicants', u'ineligible', u'laboratory', u'compliant', u'idexx', u'laboratories',
u'healthcare', u'innovation', u'practicing', u'veterinarians', u'diagnostic',
u'technology', u'idexx', u'enhance', u'veterinarians', u'efficiency', u'economically',
u'idexx', u'worldwide', u'diagnostic', u'tests', u'tests', u'quality', u'headquartered',
u'idexx', u'laboratories', u'employs', u'customers', u'qualifications', u'applicants',
u'idexx', u'interns', u'potential', u'demonstrated', u'portfolio', u'recommendation',
u'resumes', u'marketing', u'location', u'americas', u'verification', u'validation',
u'schedule', u'overtime', u'idexx', u'laboratories', u'reviews', u'idexx',
u'laboratories', u'nasdaq', u'healthcare', u'innovation', u'practicing',
u'veterinarians'], [u'location', u'duration', u'temp', u'verification', u'validation',
u'tester', u'verification', u'validation', u'middleware', u'specifically', u'testing',
u'applications', u'clinical', u'laboratory', u'regulated', u'environment',
u'responsibilities', u'complex', u'hardware', u'testing', u'clinical', u'analyzers',
u'laboratory', u'graphical', u'interfaces', u'complex', u'sample', u'sequencing',
u'protocols', u'developers', u'correction', u'tracking', u'tool', u'timely',
u'troubleshoot', u'testing', u'functional', u'manual', u'automated', u'participate',
u'ongoing', u'testing', u'coverage', u'planning', u'documentation', u'testing',
u'validation', u'corrections', u'monitor', u'implementation', u'recurrence',

u'operating', u'statistical', u'quality', u'testing', u'global', u'multi', u'teams',
u'travel', u'skills', u'concepts', u'waterfall', u'agile', u'methodologies',
u'debugging', u'skills', u'complex', u'automated', u'instrumentation', u'environment',
u'hardware', u'mechanical', u'components', u'tracking', u'lifecycle', u'management',
u'quality', u'organize', u'define', u'priorities', u'organize', u'supervision',
u'aggressive', u'deadlines', u'ambiguity', u'analyze', u'complex', u'situations',
u'concepts', u'technologies', u'verbal', u'skills', u'effectively', u'technical',
u'clinical', u'diverse', u'strategy', u'clinical', u'chemistry', u'analyzer',
u'laboratory', u'middleware', u'basic', u'automated', u'testing', u'biomedical',
u'engineering', u'technologists', u'laboratory', u'technology', u'availability',
u'click', u'attach'], [u'scientist', u'linux', u'asrc', u'scientist', u'linux', u'asrc',
u'technology', u'solutions', u'subsidiary', u'asrc', u'engineering', u'technology',
u'contracts', u'multiple', u'agencies', u'scientists', u'engineers', u'management',
u'personnel', u'allows', u'solutions', u'complex', u'aeronautics', u'aviation',
u'management', u'aviation', u'engineering', u'hughes', u'technical', u'technical',
u'aviation', u'evaluation', u'engineering', u'management', u'technical', u'terminal',
u'surveillance', u'programs', u'currently', u'scientist', u'travel',
u'responsibilities', u'develops', u'technology', u'modifies', u'technical', u'complex',
u'reviews', u'draft', u'conformity', u'completeness', u'testing', u'interface',
u'hardware', u'regression', u'impact', u'reliability', u'maintainability', u'factors',
u'standardization', u'skills', u'travel', u'programming', u'linux', u'environment',
u'cisco', u'knowledge', u'terminal', u'environment', u'clearance', u'clearance',
u'input', u'output', u'digital', u'automatic', u'terminal', u'management',
u'controller', u'termination', u'testing', u'evaluating', u'policies', u'procedure',
u'interface', u'installation', u'verification', u'certification', u'core', u'avionic',
u'programs', u'knowledge', u'procedural', u'testing', u'interfacing', u'hardware',
u'regression', u'impact', u'reliability', u'maintainability', u'factors',
u'standardization', u'missions', u'asrc', u'subsidiaries', u'affirmative', u'employers',
u'applicants', u'disability', u'veteran', u'technology', u'location', u'airport',
u'bachelor', u'schedule', u'travel', u'contributor', u'management', u'asrc',
u'reviews'], [u'technical', u'solarcity', u'niche', u'vegas', u'overview', u'resolving',
u'customer', u'clients', u'expanding', u'engineers', u'developers', u'responsibilities',
u'knowledge', u'planning', u'adapt', u'dynamic', u'environment', u'inventive',
u'creative', u'solarcity', u'lifecycle', u'responsibilities', u'technical',
u'analyzing', u'diagnosing', u'troubleshooting', u'customers', u'ticketing', u'console',
u'escalate', u'knowledge', u'engineering', u'timely', u'basic', u'phone',
u'functionality', u'customer', u'tracking', u'knowledgebase', u'rotation', u'configure',
u'deployment', u'sccm', u'technical', u'deployment', u'deploy', u'hardware',
u'solarcity', u'bachelor', u'knowledge', u'dell', u'laptops', u'analytical',
u'troubleshooting', u'solving', u'skills', u'knowledge', u'databases', u'preferably',
u'server', u'preferably', u'monitoring', u'suites', u'documentation', u'procedures',
u'knowledge', u'entries', u'verbal', u'skills', u'customer', u'skills', u'competitive',
u'solar', u'package', u'insurance', u'vacation', u'savings', u'referral',
u'eligibility', u'equity', u'performers', u'solarcity', u'affirmative', u'diversity',
u'workplace', u'applicants', u'orientation', u'disability', u'veteran',
u'careerrookie'], [u'embedded', u'exelis', u'junction', u'exelis', u'embedded',
u'acquisition', u'networking', u'capabilities', u'classified', u'customer',
u'motivated', u'develops', u'tests', u'innovative', u'solutions', u'minimal',
u'supervision', u'paced', u'environment', u'enjoys', u'assignments', u'interact',
u'multi', u'disciplined', u'challenging', u'focused', u'embedded', u'developments',
u'spanning', u'engineering', u'lifecycle', u'specification', u'enhancement',
u'applications', u'embedded', u'freescale', u'applications', u'android', u'platforms',
u'interface', u'customers', u'developers', u'refine', u'specifications',
u'architectures', u'java', u'programming', u'scripts', u'python', u'debug',
u'debugging', u'emulators', u'regression', u'revisions', u'specialized', u'setups',
u'capabilities', u'subversion', u'technical', u'documentation', u'multiple',
u'engineering', u'techexpousa', u'reviews'], [u'modeler', u'semantic', u'modeling',
u'models', u'skills', u'ontology', u'resource', u'framework', u'schema',
u'technologies', u'hadoop', u'warehouse', u'oracle', u'relational', u'artifacts',
u'models', u'dictionaries', u'models', u'interface', u'specifications',
u'documentation', u'harmonization', u'mappings', u'aligned', u'coordinate',
u'technical', u'peer', u'reviews', u'stakeholder', u'communities', u'impact',
u'domains', u'relationships', u'interdependencies', u'models', u'define', u'analyze',
u'legacy', u'models', u'corporate', u'databases', u'architectural', u'alignment',
u'customer', u'expertise', u'harmonization', u'modeling', u'modeling', u'consulting',
u'stakeholders', u'quality', u'models', u'storage', u'agile', u'specifically', u'focus',
u'modeling', u'qualifications', u'bachelors', u'accredited', u'modeler', u'encompass',
u'evaluation', u'skills', u'knowledge', u'modeling', u'techniques', u'resource',

```
u'framework', u'schema', u'technologies', u'unified', u'modeling', u'technologies',
u'schemas', u'ontologies', u'sybase', u'knowledge', u'skills', u'interpersonal',
u'skills', u'customers', u'clearance', u'applicants', u'eligibility', u'classified',
u'clearance', u'polygraph', u'techexpousa', u'solutions', u'partnership', u'solutions',
u'integration'], [u'technologies', u'junction', u'develops', u'maintains', u'enhances',
u'complex', u'diverse', u'intensive', u'analytics', u'algorithm', u'manipulation',
u'management', u'documented', u'individually', u'reviews', u'tests', u'components',
u'adherence', u'resolves', u'utilizes', u'methodologies', u'environment', u'input',
u'components', u'hardware', u'offs', u'reuse', u'cots', u'gots', u'synthesis',
u'components', u'tasks', u'individually', u'analyzes', u'modifies', u'debugs',
u'corrects', u'integrates', u'operating', u'environments', u'develops', u'queries',
u'databases', u'repositories', u'recommendations', u'improving', u'documentation',
u'develops', u'implements', u'algorithms', u'functional', u'assists', u'developing',
u'executing', u'procedures', u'components', u'reviews', u'documentation', u'solutions',
u'analyzing', u'conferring', u'users', u'engineers', u'analyzing', u'investigating',
u'areas', u'adapt', u'hardware', u'mathematical', u'models', u'predict', u'outcome',
u'implement', u'complex', u'database', u'repository', u'interfaces', u'queries',
u'bachelors', u'accredited', u'substituted', u'bachelors', u'firewalls', u'ipsec',
u'vpns', u'technology', u'administering', u'servers', u'apache', u'jboss', u'tomcat',
u'developing', u'interfaces', u'firefox', u'internet', u'explorer', u'operating',
u'mainframe', u'linux', u'solaris', u'virtual', u'scripting', u'programming',
u'oriented', u'programming', u'ajax', u'script', u'procedures', u'cobol', u'cognos',
u'fusion', u'focus', u'html', u'java', u'java', u'script', u'jquery', u'perl',
u'visual', u'basic', u'powershell', u'cots', u'cots', u'oracle', u'apex',
u'integration', u'competitive', u'package', u'bonus', u'corporate', u'equity',
u'tuition', u'reimbursement', u'referral', u'bonus', u'holidays', u'insurance',
u'flexible', u'disability', u'insurance', u'technologies', u'disability',
u'accommodation', u'recruiter', u'techexpousa']]
```

This corpus contains 10 documents. Now lets say we would like to model this with DTM. To do this we have to define the time steps each document belongs to. In this case the first 3 documents were collected at the same time, while the last 7 were collected a month later, and we wish to see how the topics change from month to month. For this we will define the `time_seq`, which contains the time slice definition.

```
In [4]: time_seq = [3, 7]  # first 3 documents are from time slice one
        #  and the other 7 are from the second time slice.
```

A simple corpus wrapper to load a premade corpus. You can use this with your own data.

```
In [5]: class DTMcorpus(corpora.textcorpus.TextCorpus):

            def get_texts(self):
                return self.input

            def __len__(self):
                return len(self.input)

        corpus = DTMcorpus(documents)
```

So now we have to generate the path to DTM executable, here I have already set an ENV variable for the DTM_HOME

```
In [6]: # path to dtm home folder
        dtm_home = os.environ.get('DTM_HOME', "dtm-master")
        # path to the binary. on my PC the executable file is dtm-master/bin/dtm
        dtm_path = os.path.join(dtm_home, 'bin', 'dtm') if dtm_home else None
        # you can also copy the path down directly. Change this variable to your DTM executable
        before running.
        dtm_path = "/home/bhargav/dtm/main"
```

That is basically all we need to be able to invoke the Training.

If `initialize_lda=True` then DTM will create a LDA model first and store it in initial-lda-ss.dat. If you already have itial-lda-ss.dat in the DTM folder then you can save time and re-use it with `initialize_lda=False`. If the file is missing then DTM wil exit with an error.

```
In [7]: model = DtmModel(dtm_path, corpus, time_seq, num_topics=2,
                         id2word=corpus.dictionary, initialize_lda=True)
```

If everything worked we should be able to print out the topics

```
In [8]: topics = model.show_topic(topicid=1, time=1, num_words=10)

In [9]: topics

Out[9]: [(0.023565028919164586, 'skills'),
         (0.02308969736545094, 'engineering'),
         (0.019616329462533579, 'idexx'),
         (0.0194313503731963, 'testing'),
         (0.01858957362093603, 'technical'),
         (0.017685337300946517, 'electrical'),
         (0.017483543705882995, 'management'),
         (0.015310984365058886, 'complex'),
         (0.014032951915032212, 'knowledge'),
         (0.012958700085355939, 'technology')]
```

1.1 Document-Topic proportions

Next, we'll attempt to find the Document-Topic proportions. We will use the gamma class variable of the model to do the same. Gamma is a matrix such that gamma[5,10] is the proportion of the 10th topic in document 5.

To find, say, the topic proportions in Document 1, we do the following:

```
In [10]: doc_number = 1
         num_topics = 2

         for i in range(0, num_topics):
             print ("Distribution of Topic %d %f" % (i, model.gamma_[doc_number, i]))

Distribution of Topic 0 0.562498
Distribution of Topic 1 0.437502
```

1.2 DIM Example

The DTM wrapper in Gensim also has the capacity to run in Document Influence Model mode. The Model is described in this paper. What it allows you to do is find the 'influence' of a certain document on a particular topic. It is primarily used in identifying the scientific impact of research papers through the capability of that document's keywords influencing a topic.

'Influence' can be naively thought of like this - if more of a particular document's words appear in subsequent evolution of a topic, that document is understood to have influenced that topic more.

To run it in this mode, we now call `DtmModel` again, but with the `model` parameter set as `fixed`.

Note that running it in this mode will also generate the DTM topics similar to running plain DTM, but with added information on document influence.

```
In [11]: model = DtmModel(dtm_path, corpus, time_seq, num_topics=2,
                          id2word=corpus.dictionary, initialize_lda=True, model='fixed')
```

The main difference between the DTM and DIM models are the addition of Influence files for each time-slice, which is interpreted with the `influences_time` variable.

To find, say, the influence of Document 2 on Topic 2 in Time-Slice 1, we do the following:

```
In [12]: document_no = 1 #document 2
         topic_no = 1 #topic number 2
         time_slice = 0 #time slice 1

         model.influences_time[time_slice][document_no][topic_no]
```

```
Out[12]: 0.0061833357763878861
```

1.3 Differences between DTM and DIM mode.

There are not too many differences in DTM and DIM apart from the Document Influence information which is generated by running it in DIM mode. The topics generated by both the models are also more or less similar.

As for running times, with smaller corpuses of less than 2000 documents, time taken for the two models is roughly the same, but for larger corpuses DIM mode takes significantly more time - usually 1.5 or 2 times as how long DTM would take.

For examples of use-cases of both, the following resources might be helpful:
Modeling Musical Influence with Topic Models
A Language-based Approach to Measuring Scholarly Impact
Studying the history of ideas using topic models