Tutorial

October 24, 2016

1 Anseri Topic Analysis Tutorial

1.1 0. Open a dataset

```
In [1]: import anseri as ai
        import numpy as np

        ai.disable_progress()  # Suppress progress notifications for a cleaner note

        d = ai.Dataset("aljazeera")
```

1.2 1. Select Data

1.2.1 Select ALL

```
In [2]: selection = ai.AllSelection()
```

1.2.2 Select by time window

```
In [3]: print(d.time_range)   # Get the unix timestamps of the min/max times of ent
        print([ai.utc.mth(x) for x in d.time_range])   # Get human-readable represe
```

```
(1300233600, 1361836800)
['Mar 16 2011', 'Feb 26 2013']
```

```
In [4]: selection = ai.TimeSelection(('Mar 1 2012', 'Jan 2 2013'))   # Time windows
        selection = ai.TimeSelection((1300233600, 1350000000))
```

1.2.3 Attribute Selection

The aljazeera dataset has no defined data attributes apart from time.

If the dataset had a column defined as an attribute, named "author", you could select documents authored by "Marwan Bishara" as follows:

```
In [5]: # selection = ai.AttributeSelection("author", ["Marwan Bishara"])
```

1.2.4 Full-text Search

```
In [6]: selection = ai.FullTextSelection("iraq war")    # full text search with fu
```

1.2.5 Select by keyword (feature) mention

```
In [7]: selection = ai.FeatureSelection("iraq")
```

1.3 2. Load data matching selection

```
In [8]: model = d.load(selection)    # Load the sparse matrix model of corpus

In [9]: len(selection.docids)    # selection.docids contain references to documents

Out[9]: 158
```

1.3.1 Load Raw Content from Database

```
In [10]: for doc in d.get_documents_by_id(list(selection.docids)[:3], fields=['titl
             print(doc.keys())
             if 'title' in doc.keys():
                 print("Title: \t\t{}".format(doc['title']))
             else:
                 print("Content: \t{}".format(doc['content']))

             print("\n\n")

DOCIDS:[1536, 12802, 1546]
['docid', 'content']
Content:         Iraq will ask the US to keep its troops in the country beyond the



['docid', 'title']
Title:                 Iraq 'to request' US troops to stay



['docid', 'content']
Content:         Hundreds of thousands of Sunni protesters have held anti-governmer



['docid', 'title']
Title:                 Iraq Sunnis rally against Shia-led government



['docid', 'content']
Content:         Twin car bombings in the northern Iraqi city of Mosul and an attac
```

```
['docid', 'title']
Title:                  Deaths in Iraq attacks
```

```
In [11]: print(doc.keys())    # Get the names of available fields in content
```

```
['docid', 'title']
```

```
In [12]: n, m = model.shape      # Get details about shape of sparse matrix represen
         print("n documents: {:,}".format(n))
         print("m features: {:,}".format(m))
```

```
n documents: 158
m features: 1,719
```

1.4 3. Get Topics

```
In [13]: # Instantiate the algorithm
         # ignore_terms injects extra stop-words at runtime

         SPCA = ai.topicmodels.TopicModelSPCA(n_topics=16,
                                              card_terms=8,
                                              card_docs=(n // 10),    # a good rule
                                              ignore_terms=["gen", "jan", "feb", "m
                                                            "aug", "sep", "oct", "n
```

```
In [14]: topics = SPCA(model, ignore_words=["gen", "jan", "feb", "mar", "apr", "may
                                                            "aug", "sep", "oct", "n
```

```
In [15]: print(topics)    # TopicModel object has a convenient representation
```

```
killed attacks people: [killed, attacks, people, iraq, officials, series, baghdad,
---
bomb city iraqi: [bomb, city, iraqi, car, police, wounded, injured, suicide]
---
minister maliki prime: [minister, maliki, prime, sunni, shia, nouri, government, pr
---
security forces medical: [security, forces, medical, gunmen, town, tuesday, attack,
---
country troops year: [country, troops, year, military, withdrawal, obama, war, pres
---
north bomber killing: [north, bomber, killing, capital, struck, mosque, south, mono
---
friday northern muslim: [friday, northern, muslim, mosul, prayers, kirkuk, blast, r
```

```
---
local wounding pilgrims: [local, wounding, pilgrims, sources, bombs, roadside, expl
---
qaeda leader thursday: [qaeda, leader, thursday, morning, province, jazeera, violer
---
official ministry interior: [official, ministry, interior, energy, oil, pipeline, t
---
hashemi court sunday: [hashemi, court, sunday, death, vice, tareq, absentia, handec
---
explosions basra southern: [explosions, basra, southern, occurred, dozens, wednesda
---
left bombings targeting: [left, bombings, targeting, worshippers, border, shootings
---
kurdish office region: [kurdish, office, region, kurdistan, party, disputed, northe
---
state department united: [state, department, united, deal, states, worth, qatar, ai
---
thousands major iraqis: [thousands, major, iraqis, highway, fallujah, demonstration
```

```
In [16]: # Topics are defined as weighted collections of words. Weights can be four
         for t in topics:
             print(t.weights)
```

```
(0.17838582753888194, 0.17637131157185096, 0.16222201222889057, 0.1568692309290613
(0.19895882078463445, 0.15281489107964669, 0.14711911586105778, 0.1317862310575131
(0.16451897624596001, 0.14644747511346373, 0.14081745902701578, 0.1407785945921067
(0.42114495144099423, 0.10142270059249049, 0.094167200348420105, 0.083484349355371
(0.1896153964005661, 0.17587849812988654, 0.12433345255586042, 0.12144872741895123,
(0.28530984514275431, 0.13616516290907085, 0.11801650618726602, 0.11729075064618, 0
(0.23764483456944852, 0.20727057321074377, 0.11369371116193044, 0.0989567103902571
(0.16999112302439093, 0.14454036601712925, 0.13212437238740499, 0.1318654028663731
(0.2719974700666371, 0.13182029456893612, 0.12419592092045272, 0.12269072840459212,
(0.38015304521401205, 0.15429737442537927, 0.099456073926220606, 0.091347543686522
(0.15351534278134857, 0.14540489563373207, 0.14108680584683536, 0.137640663292381
(0.28272367859998282, 0.18095908339019623, 0.13075113035232538, 0.1049389819050881
(0.27224513290084801, 0.21289601928381599, 0.13264367403941713, 0.0875767081790196
(0.25980970749288279, 0.19915448774086073, 0.12630855775368421, 0.1255745417986831
(0.29699126475487614, 0.15105871923047776, 0.10838821158290603, 0.1028710651139506
(0.18890520642298345, 0.13858817367271292, 0.12940533803684726, 0.1285114693641488
```

NOTE: The first set of words represent the strongest portion of the topic, covering approximately 95% of the topic strength. The words in brackets represent the total list of words defining the topic.

1.5 2. Get Documents Relevant to Topic

```
In [17]: topics.mat
```

```
Out[17]: <16x29942 sparse matrix of type '<class 'numpy.float64'>'
                with 128 stored elements in Compressed Sparse Row format>
```

1.5.1 Get Document Recommendations

```
In [18]: # Get strongest examples of a single topic
         for row in ai.topicmodels.TopicDocumentRecommendation(topics[0], model, n_
             if 'title' in row.keys():
                 print(row['title'])

         # Get strongest examples of each topic in a collection of topics
         recommendations = ai.topicmodels.TopicDocumentRecommendation(topics, model
         for i, topicdocs in enumerate(recommendations):
             print("TOPIC {}".format(i+1))
             for row in topicdocs:
                 if 'title' in row.keys():
                     print(row['title'])
```

```
SHAPE SCORE: (158, 1)
DOCIDS:[12159, 8734, 6343, 4774, 9265, 11551, 10387, 6568, 5815, 5526, 12568, 8848,
Series of deadly attacks hit Iraq
Dozens dead and wounded in Iraq bombings
Dozens dead in wave of Iraq attacks
Attacks leave many dead in Iraq
Spate of deadly attacks across Iraq
Deaths in Iraq bomb explosions
Security forces targeted in Iraq attacks
Multiple Iraq attacks leave many dead
Deaths in attacks on Iraq's Sunni districts
Many deaths in series of Iraq attacks
Deaths reported in Iraq suicide blasts
Deaths in Iraq car bomb attack
Five US troops killed in Iraq attack
Dozen killed in Iraq violence
Blast strikes Shia charity office in Iraq
Al-Qaeda group takes credit for Iraq attacks
Dozens dead in string of Iraq blasts
Trio of violent attacks strike Iraq
Iraq sees deadliest month in over two years
Attacks on Iraq's Shias leave scores dead
SHAPE SCORE: (158, 16)
DOCIDS:[12159, 6343, 4774, 8734, 5815]
DOCIDS:[8826, 12816, 1404, 2461, 465]
DOCIDS:[12166, 12071, 13056, 1469, 12631]
DOCIDS:[10387, 5342, 12226, 3211, 10414]
DOCIDS:[3956, 5058, 5114, 5084, 3877]
DOCIDS:[4719, 12832, 12568, 9283, 9572]
DOCIDS:[9572, 13206, 12954, 6892, 12802]
```

DOCIDS:[5530, 1722, 5427, 1608, 8559]
DOCIDS:[4774, 11636, 8861, 10482, 6343]
DOCIDS:[9193, 3877, 11676, 4707, 11636]
DOCIDS:[10135, 10122, 7091, 6282, 7069]
DOCIDS:[4222, 3468, 3618, 4631, 11617]
DOCIDS:[11005, 5526, 10569, 11391, 11617]
DOCIDS:[11877, 1892, 9329, 5159, 3928]
DOCIDS:[3540, 7091, 3204, 5301, 11116]
DOCIDS:[12105, 12213, 13056, 12071, 13206]
TOPIC 1
Series of deadly attacks hit Iraq
Dozens dead in wave of Iraq attacks
Attacks leave many dead in Iraq
Dozens dead and wounded in Iraq bombings
Deaths in attacks on Iraq's Sunni districts
TOPIC 2
Dozens killed in Iraq car bomb attack
Dozens killed at Iraq police headquarters
Deaths as blasts rock central Iraq city
Two deadly bombings strike Iraq
Many killed in Iraq attacks
TOPIC 3
Iraq PM warns Sunni protesters to end rallies
Iraq Sunnis block trade routes in new protest
Anti-government protests continue in Iraq
Maliki asks for patience on Iraq reforms
Several killed in clashes in Iraq's Fallujah
TOPIC 4
Security forces targeted in Iraq attacks
Gunmen kill several security guards in Iraq
Shia pilgrims killed by car bomb in Iraq
Shia pilgrims shot dead in western Iraq
Gunmen seize control of Iraq prison
TOPIC 5
Obama: All US troops to leave Iraq in 2011
Obama marks coming end of US war in Iraq
Last US combat troops leave Iraq
US forces mark end of Iraq mission
US 'abandons' plans to keep troops in Iraq
TOPIC 6
Suicide bomber strikes outside Iraq prison
Suicide bomber kills over a dozen in Iraq
Deaths reported in Iraq suicide blasts
Iraq blasts kill Kurdish security officials
Deaths in northern Iraq attacks
TOPIC 7
Deaths in northern Iraq attacks
Protests in Iraq continue amid new killings

Deadly bomb attacks rock Iraq markets
Death row inmates in Iraq prison break
Iraq Sunnis rally against Shia-led government
TOPIC 8
Shia pilgrims targeted in deadly Iraq attacks
Iraq blast hits French embassy convoy
Wave of bombings leaves scores dead in Iraq
Attackers storm government building in Iraq
Suicide bombing strikes funeral in Iraq
TOPIC 9
Attacks leave many dead in Iraq
Iraq's 'al-Qaeda chief' arrested
Iraq says al-Qaeda flowing into Syria
Many killed in string of Iraq attacks
Dozens dead in wave of Iraq attacks
TOPIC 10
Blast shuts down Iraq-Turkey oil pipeline
US 'abandons' plans to keep troops in Iraq
Iraq denies entry to Turkish minister
Iraq hit by series of fatal bombings
Iraq's 'al-Qaeda chief' arrested
TOPIC 11
Iraq sentences vice-president to death
Iraq vice-president rejects death sentence
Qatar rejects Iraq's call to extradite VP
Iraq VP rejects 'death squad charges'
Iraq demands extradition of 'fugitive' VP
TOPIC 12
Deadly triple bombing strikes Iraq oil hub
Deadly explosions hit Iraq's Karbala city
'Police chief killed' in Iraq hostage drama
Deadly triple blasts rock Iraq's south
Scores killed in Iraq blasts
TOPIC 13
Iraq rocked by wave of deadly Eid attacks
Many deaths in series of Iraq attacks
Al-Qaeda group takes credit for Iraq attacks
Syrian rebels 'seize airport near Iraq'
Scores killed in Iraq blasts
TOPIC 14
Bomb kills Kurdish security recruits in Iraq
Iraq's Kurds fear border disputes
Kurd leader warns against budget cuts by Iraq
Iraq's Maliki urges Kurds to hand over VP
Turkish troops enter Iraq after PKK attacks
TOPIC 15
Iraq makes first payment for US F-16s
Qatar rejects Iraq's call to extradite VP

```
Colin Powell regrets Iraq war intelligence
US pushes ahead with arms deal to Iraq
Iran's currency woes hurt wallets in Iraq
TOPIC 16
Iraq mass protests mount pressure on Maliki
Anti-government protests rage across Iraq
Anti-government protests continue in Iraq
Iraq Sunnis block trade routes in new protest
Protests in Iraq continue amid new killings
```

1.5.2 Note: Defining document display

Different datasets use different schemas. The following code snippet shows you how to determine the names of columns in your dataset so you can decide how to display your content.

```
In [19]: # First, get name of tables. In this example, we are interested in "Data".

         print(d.db_controller.get_table_names())
```

```
['content', 'content_content', 'content_segments', 'content_segdir', 'content_docsi
```

```
In [20]: # Example: Print title and content fields

         #for topdoc in recommendations[3]:
         #    print("""-- {title} --\n{content}\n\n""".format(**topdoc))
```

1.5.3 Export to CSV

```
In [21]: import csv
         def export_to_csv(fname, topics, topdocs, document_fields=None):
             """
             Export topics and corresponding recommended topic docs to csv file
             """
             if not document_fields:
                 document_fields = ['title', 'content']

             colnames = ['topic_index', 'topic_name'] + document_fields

             with open(fname+".csv", 'w') as csvfile:
                 writer = csv.DictWriter(csvfile, fieldnames=colnames)
                 writer.writeheader()

                 for i, (topic, docs) in enumerate(zip(topics, topdocs)):
                     record = {"topic_index": i, "topic_name": topic.name}
                     for doc in docs:
                         for df in document_fields:
                             if df in record:
```

```
                                record[df] = doc[df]
                            writer.writerow(record)
```

```
In [22]: export_to_csv('aljazeera_topics', topics, recommendations)
```

1.6 3. Subtopics

```
In [23]: selected_topic = 1
         print(topics[selected_topic])
```

```
bomb city iraqi: [bomb, city, iraqi, car, police, wounded, injured, suicide]
```

```
In [24]: # TopicSelection defines a topic "mention" based on a threshold of topic s
         # either 'abs' for a concrete value threshold, or 'quantile' to specify a
         # Then, all documents "mentioning" the given topic are selected.

         subtopic_sel = ai.topicmodels.TopicSelection(topics[selected_topic], thres
         subtopic_sel(model)    # Make the selection concrete by passing in the mod
```

```
In [25]: submodel = model[subtopic_sel]    # Data models can be sliced by a selecti
```

```
In [26]: subtopics = SPCA(submodel, ignore_words=[k for t in topics for k in t.feat
         print(subtopics)
```

```
building men provincial: [building, men, provincial, armed, council, stormed, detor
---
group members policemen: [group, members, policemen, camp, saturday, iranian, oppos
---
prison guards including: [prison, guards, including, jail, taji, hospital, injuring
---
airport army area: [airport, army, area, officers, soldiers, hold, syria, fighting]
---
tikrit bank blew: [tikrit, bank, blew, salaries, explosives, centre, scene, emergen
---
man fighters held: [man, fighters, held, months, nations, contractor, mission, appe
---
blasts hit areas: [blasts, hit, areas, wave, attacked, checkpoint, cities, hour]
---
funeral baquba place: [funeral, baquba, place, news, sheikh, village, media, quoted
---
anbar protesters led: [anbar, protesters, led, sectarian, deputy, prisoners, incide
---
convoy abu ghraib: [convoy, abu, ghraib, attempt, gmt, apparent, towns, finance]
---
market source bodies: [market, source, bodies, kilometres, confirmed, busy, receivi
---
damaged unit entrance: [damaged, unit, entrance, vehicle, main, gate, large, vehicl
---
```

```
officer west located: [officer, west, located, senior, bombers, intelligence, bodyg
---
spokesman serve civilians: [spokesman, serve, civilians, days, accused, staff, priv
---
talabani jalal suffered: [talabani, jalal, suffered, stable, rushed, treated, care,
---
headquarters attacker diyala: [headquarters, attacker, diyala, victims, early, twin
```

2 4. Regression

```
In [27]: lma = ai.linearmodels.LinearModelRS(rho=0.001)

         # What is the difference between Iraq and Iran?
         pos_sel = ai.FeatureSelection("iraq") - ai.FeatureSelection("iran")
         neg_sel = ai.FeatureSelection("iran") - ai.FeatureSelection("iraq")
         d.select(pos_sel)
         d.select(neg_sel)
         model = d.load(pos_sel + neg_sel)
         classvec = model.get_classification_vector({1.: pos_sel, -1.: neg_sel})
         print(np.min(classvec.data))
         print(np.max(classvec.data))

         selection_map = {1: pos_sel, -1: neg_sel}
         selection_map = [(s, v) for v, s in iter(selection_map.items())]

         # ignore words: iraq, iran
         model._mat = ai.data_conditioning.remove_cols(model.mat, model.feature_id_
         # Compute the solution
         linear_model = lma(model, classvec)

-1.0
1.0


In [28]: v = np.argsort(np.array(linear_model.params).ravel())[::-1][:20]   # Get t

In [29]: z = d.get_features_by_id(model.col_to_feature_id(v))

In [30]: for k in z:
             print(k)

baghdad
iraqi
killed
people
shia
maliki
```

```
police
sunni
attacks
prime
nouri
capital
city
government
series
troops
injured
bomber
northern
wounded
```

```
In [31]: # What are the image words for "iraq" in the news?
```

```
In [32]: lma = ai.linearmodels.LinearModelRS(rho=0.0001)
         pos_sel = ai.FeatureSelection("iraq")
         neg_sel = ai.AllSelection() - ai.FeatureSelection("iraq")
         d.select(pos_sel)
         d.select(neg_sel)
         model = d.load(ai.AllSelection())
         classvec = model.get_classification_vector({1.: pos_sel, -1.: neg_sel})
         selection_map = {1: pos_sel, -1: neg_sel}
         selection_map = [(s, v) for v, s in iter(selection_map.items())]
         # ignore words: iraq
         model._mat = ai.data_conditioning.remove_cols(model.mat, model.feature_id_
         # Compute the solution
         linear_model = lma(model, classvec)
         v = np.argsort(np.array(linear_model.params).ravel())[::-1][:20]   # Get t
         z = d.get_features_by_id(model.col_to_feature_id(v))
```

```
/Users/andrewgodbehere/.virtualenvs/python3dev/lib/python3.5/site-packages/ipykerne
```

```
In [33]: for k, w in zip(z, np.array(linear_model.params).ravel()[v]):
             print("{k}: {w:.2}".format(k=k, w=w))
```

```
bahrain: 0.049
political: 0.043
saleh: 0.036
shia: 0.029
ali: 0.029
gulf: 0.023
forces: 0.022
told: 0.021
saudi: 0.019
```

```
rights: 0.016
ben: 0.015
dialogue: 0.015
sunni: 0.014
human: 0.014
arabia: 0.013
abdullah: 0.012
yemen: 0.01
ennahda: 0.0099
tunisia: 0.0097
jebali: 0.0093
```

In [ ]: