Country Attribute

January 13, 2019

1 Similarities in Countries

This section is about finding similarities in Countries based on the publication it has. Two approaches were discussed amongst us - Creating a Correlation matrix based on words - Clustering based on word vectors At the moment we have decided to cluster them first

In [150]:
```
#Country
CountryVec = {}
```

In [151]:
```
for Country,words in zip(df['Country'],df['Cleaned Publications']):
    if Country in CountryVec:
        CountryVec[Country]+=words
    else:
        CountryVec[Country] = words
```

In [152]:
```
Countrydf = pd.DataFrame.from_dict(CountryVec,orient = 'index')
```

In [153]:
```
Countrydf.shape
```

Out[153]: (43, 1)

In [154]:
```
Countrydf.reset_index(inplace= True)
```

In [155]:
```
Countrydf.head()
```

Out[155]:

	index	0
0	Argentina	study uses approach terms evaluating attitudes...
1	Australia	growing number older adults finding retirement...
2	Bangladesh	paper atmospheric pressure chemical vapor depo...
3	ăBelgium	examination extensive adakite geochemical data...
4	Brazil	brasil considerar ocorrência seis grandes cerr...

In [156]:
```
Countrydf.columns=['Country','words']
```

In [157]:
```
#Country Counter
def C_counter(string):
    words = string.split(' ')
    count = Counter()
    for word in words:
        if word not in "new,using,also,two,used,may,first,one,use":
            count.update([word])
    return count.most_common(10)
```

```
In [158]: Countrydf['top10'] = Countrydf['words'].apply(lambda row:C_counter(row))

In [159]: Countrydf[['Country','top10']].head()
```

Out[159]:

	Country	top10
0	Argentina	[(los, 56), (del, 51), (que, 49), (las, 42), (...
1	Australia	[(research, 71), (design, 59), (work, 58), (pa...
2	Bangladesh	[(antenna, 14), (structure, 7), (band, 6), (pe...
3	ăBelgium	[(ttg, 16), (pecs, 16), (melting, 15), (river,...
4	Brazil	[(lung, 29), (que, 28), (foram, 25), (study, 2...

```
In [160]: Country_all_keywords = []
          for tp_lists in list(Countrydf['top10']):
              Country_all_keywords.append(tp_lists)

In [161]: vectorizer = TfidfVectorizer(min_df=5, analyzer='word', ngram_range=(1, 2), stop_wor
          vz3 = vectorizer.fit_transform(list(Countrydf['words']))

          vz3.shape
```

Out[161]: (43, 5972)

```
In [162]: svd = TruncatedSVD(n_components=50, random_state=0)
          svd_tfidf_country = svd.fit_transform(vz3)
          svd_tfidf_country.shape
```

Out[162]: (43, 43)

```
In [163]: tsne_model = TSNE(n_components=2,perplexity=10, verbose=1, random_state=0, n_iter=20
          tsne_tfidf_country = tsne_model.fit_transform(svd_tfidf_country)
          print(tsne_tfidf_country.shape)

[t-SNE] Computing 31 nearest neighbors...
[t-SNE] Indexed 43 samples in 0.168s...
[t-SNE] Computed neighbors for 43 samples in 0.552s...
[t-SNE] Computed conditional probabilities for sample 43 / 43
[t-SNE] Mean sigma: 0.314574
[t-SNE] KL divergence after 250 iterations with early exaggeration: 81.419098
[t-SNE] Error after 1600 iterations: 0.666859
(43, 2)


In [165]: tsne_tfidf_CountryDF = pd.DataFrame(tsne_tfidf_country)
          tsne_tfidf_CountryDF.columns = ['x', 'y']
          tsne_tfidf_CountryDF['Country'] = Countrydf['Country']
          tsne_tfidf_CountryDF['top10'] = Countrydf['top10']

In [166]: tsne_tfidf_CountryDF.head()
```

Out[166]:

	x	y	Country	top10
0	-163.239731	75.002831	Argentina	[(los, 56), (del, 51), (que, 49), (las, 42), (...
1	50.615387	8.885363	Australia	[(research, 71), (design, 59), (work, 58), (pa...
2	173.633179	28.864790	Bangladesh	[(antenna, 14), (structure, 7), (band, 6), (pe...
3	-53.019127	-28.799942	ăBelgium	[(ttg, 16), (pecs, 16), (melting, 15), (river,...
4	16.804359	32.778633	Brazil	[(lung, 29), (que, 28), (foram, 25), (study, 2...

In [167]:
```
from bokeh.models import ColumnDataSource, Range1d, LabelSet, Label
```

In [172]:
```
output_notebook()
FONT_SIZE = '10pt'
plot_tfidf = bp.figure(plot_width=700, plot_height=600, title="tf-idf clustering of t
    tools="pan,wheel_zoom,box_zoom,reset,hover,previewsave",
    x_axis_type=None, y_axis_type=None, min_border=1)
source = ColumnDataSource(data = tsne_tfidf_CountryDF)
labels = LabelSet(x='x', y='y', text='Country', level='glyph',text_font_size=FONT_SIZ
              x_offset=5, y_offset=5, source=source, render_mode='canvas')
plot_tfidf.scatter(x='x', y='y', source=source)
plot_tfidf.add_layout(labels)
hover = plot_tfidf.select(dict(type=HoverTool))
hover.tooltips={"words": "@top10"}

show(plot_tfidf)
```

we have discovered 3 clusters The right most one has Pakistan and Oman because their publications is focused on healthcare .They include keywords such as -Patients-stroke-hospitals,diabetes. left top cluster has alot of common spanish keywords. Surprisingly Egyptian reporters has publications on India,softwares,systems. we are yet to find out the reason.

In [173]:
```
Country_keywords_df = pd.DataFrame(index=[list(Countrydf['Country'])],
                                   columns=['keyword_{0}'.format(i) for i in range(10)],
                                   data=Country_all_keywords)
```

In [174]:
```
Country_keywords_df[Country_keywords_df.index!=0]
```

Out[174]:

	keyword_0	keyword_1	keyword_2	\
Argentina	(los, 56)	(del, 51)	(que, 49)	
Australia	(research, 71)	(design, 59)	(work, 58)	
Bangladesh	(antenna, 14)	(structure, 7)	(band, 6)	
ăBelgium	(ttg, 16)	(pecs, 16)	(melting, 15)	
Brazil	(lung, 29)	(que, 28)	(foram, 25)	
Canada	(data, 93)	(research, 87)	(model, 87)	

Chile	(los, 23)	(study, 20)	(research, 16)
Paraguay	(species, 16)	(gnathostoma, 9)	(mexico, 8)
Columbia	(los, 19)	(que, 17)	(para, 17)
Costa Rica	(data, 26)	(species, 17)	(analysis, 14)
Costa Rica	(recognition, 24)	(language, 23)	(systems, 16)
Egypt	(india, 30)	(system, 29)	(software, 28)
El Salvador	(system, 41)	(los, 28)	(las, 27)
England	(study, 67)	(data, 49)	(model, 47)
Ethiopia	(object, 23)	(family, 22)	(pose, 19)
France	(flow, 13)	(inclination, 13)	(number, 13)
Germany	(time, 30)	(study, 27)	(response, 27)
Ghana	(hiv, 18)	(study, 10)	(health, 10)
Greece	(control, 14)	(simulation, 13)	(channel, 12)
Guinea	(newspapers, 5)	(named, 5)	(entities, 5)
India	(study, 52)	(results, 32)	(model, 31)
Ireland	(management, 27)	(knowledge, 26)	(students, 24)
Kenya	(amf, 21)	(soil, 17)	(study, 9)
Nicaragua	(que, 17)	(los, 14)	(del, 12)
Northern Ireland	(children, 26)	(years, 13)	(age, 12)
Oman	(patients, 14)	(plant, 11)	(different, 9)
Pakistan	(stroke, 31)	(patients, 27)	(hospital, 17)
palestine	(response, 14)	(task, 9)	(condition, 8)
Panama	(bioweapons, 34)	(pedestrian, 18)	(model, 17)
Peru	(los, 19)	(network, 15)	(services, 10)
Scotland	(behaviour, 32)	(nicotine, 32)	(data, 31)
United Kingdom	(study, 146)	(patients, 132)	(data, 120)
South Africa	(model, 63)	(data, 57)	(process, 55)
Sudan	(data, 33)	(methods, 28)	(results, 26)
Switzerland	(management, 29)	(model, 21)	(development, 18)
syria	(patients, 35)	(laa, 15)	(thrombus, 15)
Tanzania	(games, 58)	(paper, 26)	(study, 26)
Turkey	(follicles, 28)	(oocytes, 25)	(primordial, 23)
Uganda	(wave, 27)	(swidden, 26)	(energy, 22)
USA	(study, 755)	(data, 672)	(research, 489)
Wales	(patients, 42)	(species, 31)	(data, 30)
Zimbabwe	(results, 29)	(model, 28)	(presented, 27)

	keyword_3	keyword_4	keyword_5 \
Argentina	(las, 42)	(para, 30)	(con, 24)
Australia	(paper, 58)	(study, 56)	(health, 54)
Bangladesh	(performance, 6)	(ebg, 6)	(algorithm, 5)
ăBelgium	(river, 14)	(data, 13)	(sites, 13)
Brazil	(study, 25)	(markers, 25)	(uma, 21)
Canada	(social, 81)	(paper, 79)	(study, 78)
Chile	(mares, 16)	(social, 15)	(para, 15)
Paraguay	(larvae, 6)	(fish, 6)	(parasite, 6)
Columbia	(del, 17)	(las, 14)	(health, 12)
Costa Rica	(paper, 13)	(study, 13)	(network, 13)

Costa Rica	(system, 15)	(dnn, 15)	(based, 14)
Egypt	(flow, 26)	(equation, 25)	(feedback, 22)
El Salvador	(information, 26)	(systems, 26)	(que, 22)
England	(results, 44)	(however, 38)	(patients, 37)
Ethiopia	(dna, 17)	(reunification, 15)	(image, 14)
France	(enclosure, 11)	(numerical, 10)	(angles, 9)
Germany	(music, 27)	(mean, 20)	(practice, 20)
Ghana	(conocimientos, 8)	(del, 8)	(women, 7)
Greece	(provide, 11)	(models, 10)	(packet, 10)
Guinea	(ner, 4)	(newspaper, 4)	(europeana, 3)
India	(des, 30)	(high, 29)	(speech, 26)
Ireland	(research, 22)	(writing, 20)	(history, 18)
Kenya	(maize, 9)	(tree, 7)	(practices, 7)
Nicaragua	(las, 11)	(por, 8)	(con, 8)
Northern Ireland	(born, 10)	(spirometry, 10)	(lung, 10)
Oman	(stroke, 9)	(water, 8)	(years, 8)
Pakistan	(study, 14)	(years, 10)	(diabetics, 10)
palestine	(showed, 7)	(results, 7)	(head, 6)
Panama	(nuclear, 17)	(pedestrians, 16)	(states, 16)
Peru	(embedded, 9)	(libro, 8)	(paper, 7)
Scotland	(study, 28)	(energy, 22)	(many, 21)
United Kingdom	(results, 93)	(studies, 83)	(species, 67)
South Africa	(study, 44)	(training, 43)	(information, 41)
Sudan	(study, 24)	(host, 24)	(based, 23)
Switzerland	(stakeholders, 18)	(cell, 17)	(marine, 16)
syria	(data, 13)	(results, 13)	(heat, 12)
Tanzania	(different, 24)	(systems, 24)	(theory, 23)
Turkey	(results, 13)	(uncertainty, 13)	(activation, 12)
Uganda	(reo, 21)	(data, 19)	(systems, 19)
USA	(results, 489)	(analysis, 483)	(patients, 438)
Wales	(model, 30)	(results, 27)	(study, 26)
Zimbabwe	(different, 24)	(systems, 23)	(system, 23)

	keyword_6	keyword_7	keyword_8 \
Argentina	(por, 23)	(una, 22)	(derecho, 20)
Australia	(data, 53)	(social, 49)	(surface, 43)
Bangladesh	(design, 4)	(operating, 4)	(array, 4)
ãBelgium	(western, 12)	(birds, 12)	(infusion, 12)
Brazil	(por, 21)	(results, 21)	(cell, 21)
Canada	(play, 72)	(results, 71)	(groups, 65)
Chile	(del, 14)	(media, 13)	(sperm, 13)
Paraguay	(oaxaca, 5)	(human, 4)	(sequences, 4)
Columbia	(care, 10)	(una, 10)	(datos, 10)
Costa Rica	(large, 12)	(risk, 12)	(patients, 12)
Costa Rica	(neural, 12)	(deep, 11)	(speech, 11)
Egypt	(model, 20)	(africa, 19)	(development, 19)
El Salvador	(model, 21)	(data, 21)	(aer, 21)
England	(rock, 36)	(research, 34)	(models, 30)

Ethiopia	(objects, 13)	(approach, 11)	(camera, 11)
France	(heat, 7)	(transfer, 7)	(lattice, 7)
Germany	(performance, 19)	(heart, 16)	(results, 15)
Ghana	(vih, 7)	(among, 6)	(prevention, 6)
Greece	(propagation, 10)	(radio, 8)	(communication, 8)
Guinea	(recognition, 3)	(articles, 3)	(wikidata, 3)
India	(water, 26)	(method, 24)	(methods, 24)
Ireland	(study, 15)	(learning, 15)	(atlantic, 13)
Kenya	(agricultural, 6)	(species, 6)	(farms, 5)
Nicaragua	(como, 6)	(una, 6)	(ciudad, 5)
Northern Ireland	(function, 10)	(results, 8)	(vitamin, 8)
Oman	(field, 7)	(species, 7)	(clinical, 7)
Pakistan	(diabetic, 9)	(department, 7)	(neurology, 7)
palestine	(stop, 6)	(required, 6)	(responses, 6)
Panama	(iran, 16)	(program, 14)	(development, 13)
Peru	(del, 7)	(que, 7)	(devices, 7)
Scotland	(behavior, 20)	(different, 20)	(results, 18)
United Kingdom	(human, 65)	(analysis, 65)	(high, 62)
South Africa	(number, 39)	(analysis, 36)	(approach, 35)
Sudan	(error, 22)	(sequencing, 22)	(studies, 21)
Switzerland	(service, 15)	(dopamine, 14)	(process, 13)
syria	(equations, 12)	(disease, 11)	(risk, 11)
Tanzania	(learning, 23)	(research, 22)	(equations, 22)
Turkey	(exchange, 12)	(however, 11)	(rate, 11)
Uganda	(model, 18)	(species, 14)	(land, 14)
USA	(model, 398)	(studies, 361)	(social, 339)
Wales	(control, 24)	(noise, 23)	(plasma, 22)
Zimbabwe	(protein, 23)	(coq, 23)	(study, 22)

	keyword_9
Argentina	(como, 19)
Australia	(time, 42)
Bangladesh	(proposed, 4)
ăBelgium	(increased, 11)
Brazil	(species, 20)
Canada	(analysis, 64)
Chile	(information, 13)
Paraguay	(hosts, 4)
Columbia	(proceso, 10)
Costa Rica	(brain, 12)
Costa Rica	(lid, 10)
Egypt	(pressure, 19)
El Salvador	(paper, 20)
England	(sound, 29)
Ethiopia	(immigration, 10)
France	(applied, 7)
Germany	(movement, 15)
Ghana	(chile, 6)

Greece	(vehicular, 8)
Guinea	(historical, 2)
India	(present, 24)
Ireland	(including, 12)
Kenya	(trees, 5)
Nicaragua	(áreas, 5)
Northern Ireland	(respiratory, 7)
Oman	(study, 6)
Pakistan	(risk, 7)
palestine	(fmri, 5)
Panama	(knowledge, 13)
Peru	(web, 7)
Scotland	(performance, 18)
United Kingdom	(model, 62)
South Africa	(results, 33)
Sudan	(high, 21)
Switzerland	(fishery, 12)
syria	(effects, 11)
Tanzania	(field, 21)
Turkey	(evidence, 10)
Uganda	(language, 13)
USA	(however, 338)
Wales	(information, 21)
Zimbabwe	(increased, 20)