Problem A 4.2

Imports required for the problem

In [ ]:
```
using SparseArrays
include("wikipedia_corpus.jl")
include("kmeans.jl")
using Main.Kmeans
using Plots
```

```
(process:6212): GLib-GIO-WARNING **: 19:21:27.688: Unexpectedly, UWP app `A
cerIncorporated.PredatorSenseV30_3.0.3136.0_x64__48frkmn4z8aw4' (AUMId `Ace
rIncorporated.PredatorSenseV30_48frkmn4z8aw4!CentenialConvert') supports 1
extensions but has no verbs
```

(a) For each of k = 2, k = 5, and k = 10 run k-means twice, and plot J (vertically) versus iteration (horizontally) for the two runs on the same plot. Create your plot by passing a vector containing the value of J at each iteration to PyPlot's plot function. Comment briefly on your results

Running K-Means with k = 2:

Run 1:

In [ ]:
```
centroids, labels, j_hist_run_k2_1 = kmeans(article_histograms, 2)
```

Out[ ]:
```
(Vector{T} where T[[0.009868666739802369, 0.006359422882049034, 0.003423355
4719142983, 0.00427412914822535, 0.004187575440154912, 0.00678512327098219
3, 0.006116560611510877, 0.005315740285404562, 0.01351484690122658, 0.00420
9607428981492  …  0.01396254077139447, 0.03455377322682018, 0.0041355730618
99561, 0.008911340763341139, 0.013793862753155263, 0.020345048693745354, 0.
0034076559129716665, 0.008668468102706316, 0.01163922894498386, 0.000301884
8538328947], [0.007336785611278229, 0.007642731990440537, 0.007049150557867
311, 0.004459548067169194, 0.0023014957730097306, 0.0032558348590759683, 0.
0018432824417179 03, 0.002237238900618065, 0.002544894761246559, 0.004321836
782606721  …  0.0018742654422459679, 0.001558796858750484, 0.00428813860767
7366, 0.002340159264938871, 0.0, 0.006972120753976558, 0.00523343211543161
3, 0.0018030373998146776, 0.0014567837920983332, 0.004594761242696075]],
[1, 2, 2, 1, 2, 2, 2, 2, 2, 2  …  2, 1, 1, 1, 1, 1, 1, 1, 1, 2], [0.9121003
173639827, 0.9010867111632551, 0.8995995732564892, 0.8993867283991885, 0.89
938867283991885])
```

Run 2:

In [ ]:
```
centroids, labels, j_hist_run_k2_2 = kmeans(article_histograms, 2)
```

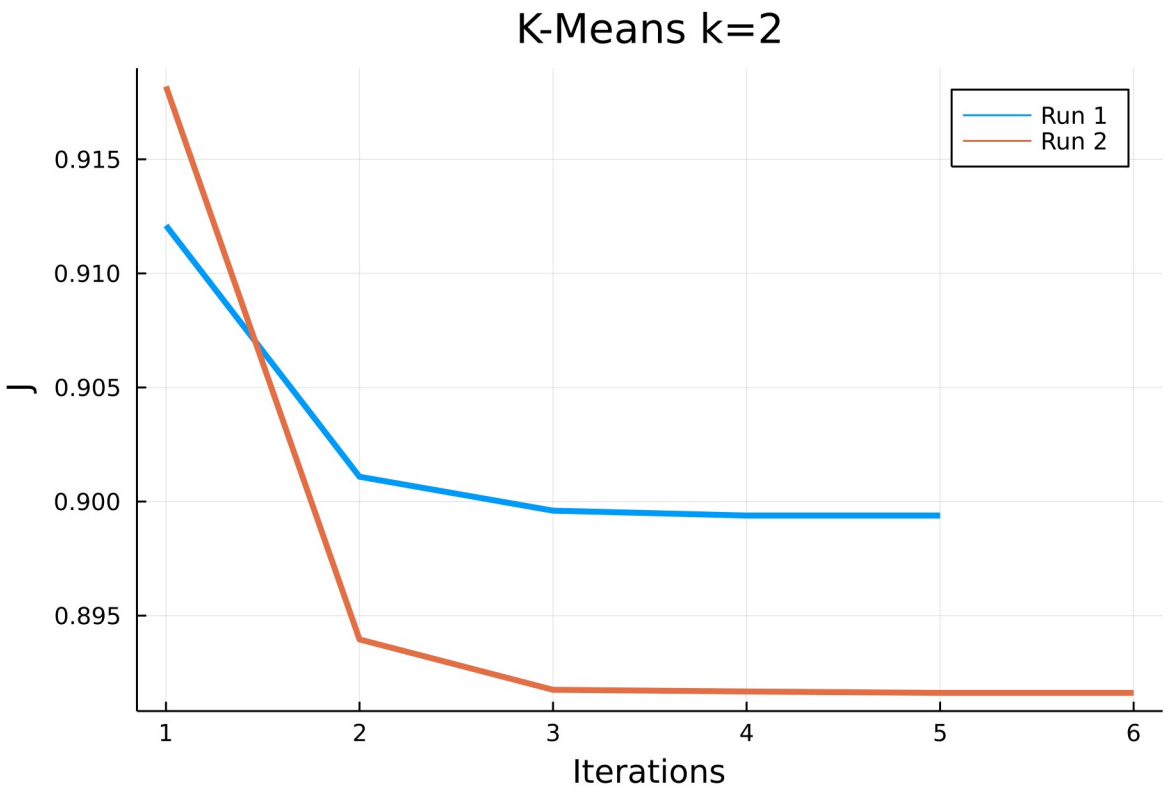
Out[ ]:
```
(Vector{T} where T[[0.00870287181245273, 0.007529761733133819, 0.0059849035
71802316, 0.004798646282069548, 0.0031946254837661356, 0.00519571090756768
2, 0.0033471443004977723, 0.002780362805510046, 0.0024406089505311372, 0.00
5069903786655501  …  0.005333847745765909, 0.006150318551711911, 0.00468506
8520808138, 0.0014871923447666816, 0.007147728881180455, 0.0129165243320655
47, 0.0015921026551624997, 0.004784401898788318, 0.00234158180306209, 0.003
6744102664050004], [0.007187979166195251, 0.006124684718576126, 0.004809071
772062999, 0.0032628060166982496, 0.0025330525941115, 0.002950411712690125,
0.0037970837220282494, 0.005130512635485875, 0.0184638625401855, 0.00210472
56925566258  …  0.00958620645160275, 0.03594995352760588, 0.002979175443834
3743, 0.014049751930635623, 0.0, 0.009681433228402373, 0.01264535704266625
2, 0.00338752377925775, 0.013533573604809872, 0.0010083775733664999]], [2,
1, 1, 2, 1, 1, 1, 1, 1, 1  …  1, 2, 1, 1, 1, 1, 1, 1, 1, 2], [0.91819645560
```

```
40284, 0.8939558966871416, 0.8917484783644691, 0.891674423577976, 0.8916143
```

Plotting the histograms for k=2

In [ ]:
```
plot([j_hist_run_k2_1, j_hist_run_k2_2], title = "K-Means k=2", label = ["R
```

Out[ ]:



Running K-Means with k = 5:

Run 1:

In [ ]:
```
centroids, labels, j_hist_run_k5_1 = kmeans(article_histograms, 5)
```

Out[ ]:
```
(Vector{T} where T[[0.010343710541156974, 0.007533285393684393, 0.002857457
897858788, 0.00536550170648606, 0.005550730468476969, 0.009099638052387879,
0.007170981499404696, 0.004070001502000909, 0.0004905401909043939, 0.005770
36186611803  …  0.012709093978610303, 0.01515216557091682, 0.00690085596122
106, 0.0011618455654507575, 0.02382576293726818, 0.02578823995364182, 0.000
24395367288878786, 0.011798841033421667, 0.0054758767774687874, 0.000435388
8714909091], [0.008794259383293696, 0.004663630820796087, 0.004232448150109
131, 0.002894078484121956, 0.002368043321478478, 0.003680259990178261, 0.00
48244562735913045, 0.007288753077545217, 0.03211106528727913, 0.00199350743
53843478  …  0.016368031420667176, 0.06374376494064782, 0.00017981601325434
783, 0.019502413088988698, 0.0, 0.013126053268880649, 0.003144625124579782
5, 0.0045539533804932605, 0.020431888840254344, 0.0001234610395336956 6],
[0.002564537513901077, 0.005964076683327539, 0.004920387558878614, 0.004460
43477782477, 0.0008684474346567692, 0.0030564992460903077, 0.00235774710976
3077, 0.0025513415232819997, 0.002771297842664461, 0.0038073669625247695  …
0.0016843222923247692, 0.0024402571787266155, 0.004696023032057539, 0.00203
9346449969077, 0.0, 0.002163947281235846, 0.004168123296543076, 0.002029934
9155987693, 0.0016034764428296921, 0.004678117080952922], [0.00487149719040
3142, 0.00786993599227057, 0.0054295179102862865, 0.0036541963161785714, 0.
0026677548992597429, 0.0019068850704859997, 0.0023383345479159995, 0.0021473
81979194, 0.0, 0.002190763239338571  …  0.0003990591650722857, 0.0016429362
839334289, 0.006573214254201427, 0.006481975781639142, 0.0, 0.0048776059401
05142, 0.024770737362360858, 0.0017577156239408573, 0.004080542906659715,
0.002142599944307714], [0.01210513596158716, 0.008769156373749773, 0.009184
```

```
78833194, 0.004677850100779659, 0.0031820490029134093, 0.00390690724103886
3, 0.0012481068751613636, 0.0020138922621229546, 0.0036866358720905683, 0.0
05534728378821705  …  0.0025586971872626134, 0.0009169404781804547, 0.00306
8375955152955, 0.00134026124182875, 0.0, 0.011351760714019662, 0.0007185639
482930683, 0.0016124902274281818, 0.0005626609156453408, 0.0054040566230540
91]], [2, 5, 5, 2, 3, 3, 3, 5, 3, 5  …  3, 2, 1, 1, 1, 1, 1, 1, 1, 4], [0.8
672836877973623, 0.8129869095113191, 0.7912010900902314, 0.779046987555712,
0.7754105407621079, 0.7751865806457023, 0.775086991514741, 0.77508699151474
```

Run 2:

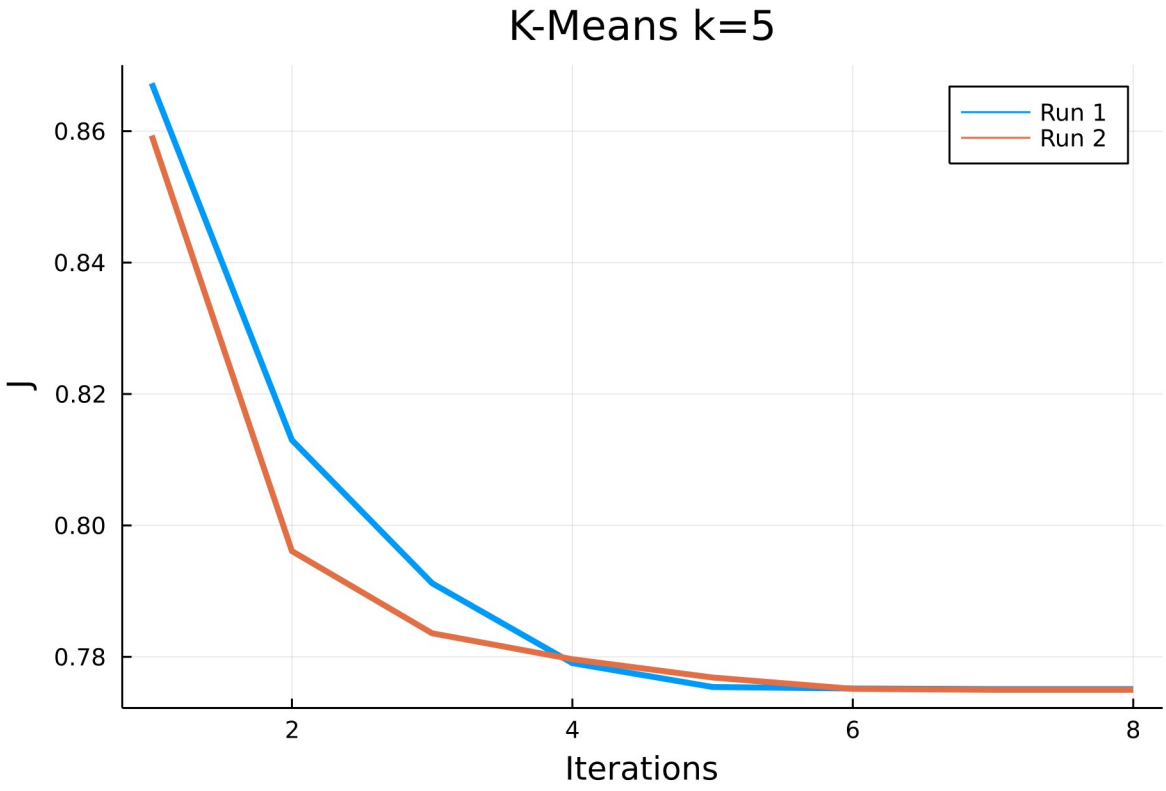
In [ ]:
```
centroids, labels, j_hist_run_k5_2 = kmeans(article_histograms, 5)
```

Out[ ]:
```
(Vector{T} where T[[0.010280063080236311, 0.007627975546130306, 0.002876292
0297695383, 0.005448047886585845, 0.0056058000635652305, 0.009187259580656,
0.007251370606073692, 0.004132616909724, 0.0004980869630721538, 0.005859136
6640583075  …  0.012775731399360002, 0.015187247837806306, 0.00697919685528
3692, 0.0010183075527361537, 0.024192313136303076, 0.02616551751978369, 0.0
0024770680631784614, 0.011921987743953538, 0.005560121035583692, 0.00044208
716182153845], [0.008794259383293696, 0.004663630820796087, 0.0042324481501
09131, 0.002894078484121956, 0.002368043321478478, 0.003680259990178261, 0.
0048244562735913045, 0.007288753077545217, 0.03211106528727913, 0.001993507
4353843478  …  0.016368031420667176, 0.06374376494064782, 0.000179816013254
34783, 0.019502413088988698, 0.0, 0.013126053268880649, 0.00314462512457978
25, 0.0045539533804932605, 0.020431888840254344, 0.00012346103953369566],
[0.003008054726106216, 0.0059695593046236485, 0.004507945192052026, 0.00401
1385574651892, 0.000869668356072973, 0.002779900967527027, 0.00239341349088
14866, 0.0022693087952744594, 0.0027161428240567564, 0.004261146244111216
…  0.0016435297584362163, 0.002444597501203108, 0.0050456739341812155, 0.00
1963385740062838, 0.0, 0.0024106652838789193, 0.0037397308221127027, 0.0018
747598092148648, 0.0014380981957155404, 0.004202072247537838], [0.012335321
06978771, 0.009137775976821446, 0.009654020308146868, 0.004910713007243734,
0.003348372479351928, 0.0040984560882560235, 0.00129825617889, 0.0021100104
485522893, 0.003657394548638072, 0.005163957049878194  …  0.002667503144407
47, 0.0008893581828034939, 0.003062247307378795, 0.0013939973512791567, 0.
0, 0.011770099465696026, 0.0007335769751126506, 0.0017096281929359037, 0.00
057013087804542 17, 0.005729602202756145], [0.005328200052003436, 0.00737481
6746940937, 0.005777540654170938, 0.0039076321365, 0.0028088888406909377,
0.0020856555458440623, 0.00193770327245875, 0.0023486990397434374, 0.0, 0.0
02102450189785937  …  0.0004364709617978125, 0.00171766008113, 0.0056119806
32485938, 0.007089661011167812, 0.0, 0.004883919859760937, 0.02698470248865
7187, 0.0018289967287125001, 0.004463093804159063, 0.0021286019373409374]],
[2, 4, 4, 2, 3, 3, 3, 4, 3, 4  …  3, 2, 1, 1, 1, 1, 1, 1, 1, 5], [0.8593358
206036759, 0.7960915032497563, 0.7835834958055152, 0.77963537667526, 0.7768
547141755753, 0.7751261775259944, 0.7749878638103469, 0.7749878638103469])
```

Plotting the histograms for k=5

In [ ]:
```
plot([j_hist_run_k5_1, j_hist_run_k5_2], title = "K-Means k=5", label = ["F
```

Out[ ]:



Running K-Means with k = 10:

Run 1:

In [ ]:
```
centroids, labels, j_hist_run_k10_1 = kmeans(article_histograms, 10)
```

Out[ ]:
```
(Vector{T} where T[[0.005161358808247307, 0.008077268153905769, 0.008899361
60191577, 0.005249683997725, 0.00727968736800423, 0.0036461206275115385, 0.
0010047552396416923, 0.0003214397858492308, 0.007128910172188077, 0.00329807
11417999998  …  0.0037930339001661538, 0.0, 0.00524575024119, 0.00314359850
7952308, 0.0, 0.009304963760036923, 0.0002962986043165385, 0.00090234040951
15385, 0.00033706046590153846, 0.0033296465873269235], [0.00885734444018054
1, 0.005705610175664325, 0.005082301499589459, 0.0034767783309397294, 0.002
679732468201351, 0.004121453240402162, 0.005997972664464865, 0.009061693015
326486, 0.036262473101647034, 0.0023630701056289185  …  0.0170931787579105
4, 0.06804292971390001, 0.0002235550435054054, 0.021936908051051893, 0.0,
0.01338969089627054, 0.003501788942949729, 0.005152889058394324, 0.02124973
0807669186, 0.00015349210320405405], [0.00961369343881, 0.00118047447463666
67, 0.0013927112004066666, 0.0018665237448708336, 0.00031914765323333335,
0.002107163078603333, 0.002025483765441667, 0.0, 0.011231669399025, 0.00121
44775514049998  …  0.010662750976511667, 0.03069856100938417, 0.00500633604
1645, 0.004517759800163334, 0.0, 0.010139554143544166, 0.002096799660981666
4, 0.0006126023925158334, 0.007992098375211667, 0.0009655691657341666], [0.
04896195180135625, 0.00797211870637375, 0.026911533225898752, 0.00188414559
89849999, 0.0, 0.0017731213101212502, 0.0014418275007875, 0.0, 0.0027173078
07875, 0.0050917221321962495  …  0.0, 0.0, 0.0, 0.0, 0.0, 0.027228057304046
25, 0.0012626533060125, 0.0, 0.00059004754311375, 0.0099813942683125], [0.0
04255563133203846, 0.007204953344295384, 0.009714747494036922, 0.0029372098
827884613, 0.0, 0.002128280264612308, 0.0013132634221723076, 0.000982710962
7423077, 0.0015280399646784616, 0.009217192757092306  …  0.0046327431659507
695, 0.003600192779934616, 0.0, 0.0005394817419492308, 0.0, 0.0004412436981
6615387, 0.0008228928997615385, 0.00141888622825, 0.0002563627621492307 7,
0.020672946432723074], [0.005166739444366969, 0.008199868220894241, 0.00560
24636646506065, 0.0037892190414545454, 0.002723770997033636 6, 0.002022453866
26366665, 0.0018789849914751515, 0.0022775263415693936, 0.0, 0.002038739577
974242  …  0.0004232445690160606, 0.0016656097756412123, 0.0054419206133196
97, 0.006874822798708181, 0.0, 0.00473592228825303, 0.02616698423142515, 0.
```

001773572585418182, 0.004327848537366364, 0.002064098848330606], [0.0098523
71625865336, 0.008050752379889165, 0.002787088957079, 0.005228948586394666,
0.005879801448811501, 0.009542234747407001, 0.007402293016193166, 0.0037101
76908257167, 0.0005395942099948333, 0.0060816963141051665 … 0.01236118046
8841668, 0.0160629800883135, 0.0072493252200823335, 0.0011031665154641666,
0.007719151066245, 0.027578521507861997, 0.00026834904017766667, 0.01274055
2495796333, 0.006023464455215666, 0.00047892775864], [0.001402901386458125,
0.005740482912757708, 0.004923381384738124, 0.005283475995887501, 0.0006743
707710429167, 0.0025234869749470837, 0.002502964579047083, 0.00325463603785
9583, 0.0021310501894014583, 0.0029429273369616666 … 0.001288569919572916
8, 0.0021991304224925, 0.006087618071481042, 0.0015570943926095833, 0.0, 0.
0022648711713935416, 0.0007262097740916668, 0.002608561118035625, 0.0009892
457832060416, 0.005206125026955], [0.010228119744383637, 0.0027126460576018
18, 0.005219528125609091, 0.005117418045758181, 0.0036270241546009093, 0.00
8688156637102727, 0.003139287937952727, 0.0024043549546436363, 0.0056720191
3550909, 0.000885194686289091 … 0.010130618281042727, 0.006511413110454545
5, 0.0013109358157536366, 0.0042329742109, 0.10085011726227272, 0.004847171
426893637, 0.0, 0.0026655467228454545, 0.009098375685781819, 0.0], [0.01007
456768022404, 0.009514917171597694, 0.005342994203558461, 0.004370371406564
231, 0.001889816859339231, 0.0040131865286294225, 0.001485494048036154, 0.0
035628094509751 92, 0.0021834330023111537, 0.0067743190906584625 … 0.00177
2534533369423, 0.0019300960294119235, 0.0027527534587555766, 0.001580279625
1184613, 0.0, 0.010779570889726729, 0.005080601378207886, 0.002041592921252
5, 0.0009053691754296151, 0.0017280620927213465]], [2, 10, 5, 2, 8, 8, 8, 1
0, 8, 1 … 8, 2, 7, 7, 7, 7, 7, 7, 9, 6], [0.8337190605769048, 0.755681578
4598209, 0.7467967675579711, 0.7429901228854263, 0.7424754478863873, 0.7419
496918218049, 0.7419496918218049])

Run 2:

In [ ]:
```
centroids, labels, j_hist_run_k10_2 = kmeans(article_histograms, 10)
```

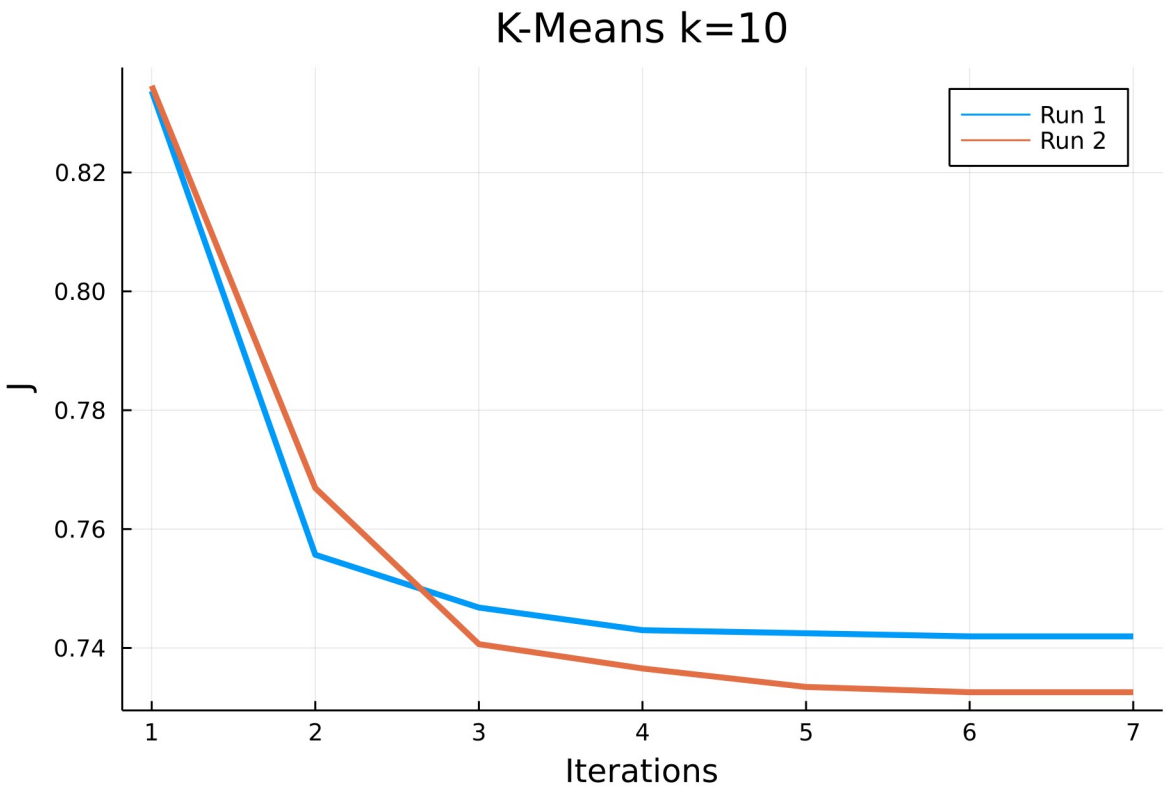
Out[ ]:
(Vector{T} where T[[0.011136536240331283, 0.008913397205784872, 0.002924059
505264359, 0.0026618083511435898, 0.005284553563182051, 0.01112151823878102
8, 0.004033321745971539, 0.0038673917069402563, 0.0006915205020651282, 0.00
6084966840176154 … 0.013712875643580253, 0.019991805749951536, 0.00591545
2402443076, 0.0007221461377787179, 0.011875617024992308, 0.0258099757230520
48, 9.774789250666666e-5, 0.01738482777653231, 0.003608595563505128, 0.0],
[0.008794259383293696, 0.004663630820796087, 0.004232448150109131, 0.002894
078484121956, 0.002368043321478478, 0.003680259990178261, 0.004824456273591
3045, 0.007288753077545217, 0.03211106528727913, 0.0019935074353843478 …
0.016368031420667176, 0.06374376494064782, 0.00017981601325434783, 0.019502
413088988698, 0.0, 0.013126053268880649, 0.0031446251245797825, 0.004553953
3804932605, 0.020431888840254344, 0.0001234610395336 9566], [0.0090479851291
725, 0.005975793324720834, 0.00303836088457125, 0.0101685256327975, 0.00631
03447931191665, 0.006563918053213334, 0.0128114154234625, 0.004349876050218
749, 0.00022526470913125, 0.005732009659037917 … 0.012075579126040417, 0.
0081807912242125, 0.007983553881935833, 0.00158442881477, 0.046222970411875
, 0.027832408475295833, 0.0005120322751208333, 0.0034904960339175, 0.00860
1664744135, 0.0011973193966], [0.002158511784461522, 0.005834291792043042,
0.004230431363371521, 0.005825414740800216, 0.001114545319907826, 0.0041876
08192490435, 0.0029802543789254348, 0.0036051565002897827, 0.00347580052068
41297, 0.003529760687027391 … 0.001961440307191739, 0.002683948659228695
3, 0.006545340467645217, 0.0018443787353858696, 0.0, 0.002832301061305, 0.0
0046599288687173917, 0.0027219768188197825, 0.001684547930684565, 0.0051419
03610849782], [0.01061420457436, 0.006831154260065237 4, 0.01002361194122333
2, 0.0026627231735090475, 0.003912514096962381, 0.0034676854052914288, 0.00
1189209480542381, 0.0007348670798909524, 0.0016019799720819049, 0.007620730
305559525 … 0.002991260126325714, 0.0010951039481190475, 0.00057056006538
38095, 0.0003840095627985 7146, 0.0, 0.0058680934588033334, 0.0013572665646
3, 0.001357267722914762, 0.0008007937317280952, 0.016398869560557146], [0.0
0248279569447, 0.011621761096224233, 0.005242886324915769, 0.004307314500308
8847, 0.004559958798988462, 0.0029025370571176923, 0.0006748647512646154,

```
0.001502847182723077, 0.004851446598295, 0.004335523138639615  …  0.0016623
07133418846, 0.0, 0.0028680966239138465, 0.0010946778139461538, 0.0, 0.0093
47813368722306, 0.0, 0.00010158123828538462, 0.0, 0.0020402404921303846],
[0.005500077473035805, 0.007612714061358387, 0.005963912933337743, 0.004033
6847860645165, 0.0028994981581325807, 0.0021529347570003222, 0.002000209829
634839, 0.0024244635248964514, 0.0, 0.0021702711636499995  …  0.00045055067
024290325, 0.001773068470843871, 0.005640922247345162, 0.00731835975346354
8, 0.0, 0.005041465661688709, 0.027855176762484838, 0.0018879966231870968,
0.004434389132799033, 0.002197266515964839], [0.0039362498508588, 0.0073819
753090868, 0.005339643777000001, 0.0032832177551148003, 0.000718771396332,
0.0066323967183184, 0.0017955151737496, 0.0026543966969400003, 0.0048074891
547716, 0.0011777583710504  …  0.0039268540390624, 0.0015902814924052, 0.00
33849468423176, 0.003167038246992, 0.0, 0.0094591467064968, 0.0, 0.00344170
55455508, 0.0, 0.0003682783822716], [0.004591069946874737, 0.00591193091784
52635, 0.006995007452348948, 0.0013609734474405264, 0.0006648483557536841,
0.0004896245720010526, 0.0015052143621931578, 0.0006555141316589474, 0.0010
015972198163157, 0.007389714405150002  …  0.0013027423044957894, 0.00211338
4145468947, 0.005352675189796315, 0.0029252728215584207, 0.0, 0.00334811799
1994737, 0.0122114426646, 0.0010492353487463158, 0.002419026926507368, 0.00
076390216306473 68], [0.03003374525223652, 0.008438528423238261, 0.014220202
427818263, 0.007255839250610435, 0.002891708742395652, 0.001554479843416956
6, 0.001666734610071739, 0.0024911476117143476, 0.0019858576326304347, 0.00
630081474719043 55  …  0.0012888825186682 61, 0.0017173253382982 61, 0.0038973
30964123043, 0.00021169301422999998, 0.0, 0.01673249566467913, 0.0024204908
039113042, 0.001367287080282174, 0.0014373204205486959, 0.00560170212284173
95]], [2, 10, 10, 2, 4, 4, 4, 6, 4, 6  …  4, 2, 1, 3, 1, 3, 1, 1, 3, 7],
[0.834594738147752, 0.7669003111093167, 0.7406400992636203, 0.7365493885718
291, 0.7334425489183125, 0.732556608476057, 0.732556608476057])
```

Plotting the histograms for k=10

In [ ]:
```
plot([j_hist_run_k10_1, j_hist_run_k10_2], title = "K-Means k=10", label =
```

Out[ ]:



With k=2 the graphs varied a lot. In case of k=5, after 4 iterations, the graphs were fairly similar. In case of k=10, the graphs were not as similar as in the case of k=5. So k=5 should be an optimum number of clusters in the scenario provided.

b)

Choose a value of k from part (a) and investigate your results by looking at the words and article titles associated with each centroid. Feel free to visit Wikipedia if an article's content is unclear from its title. Give a short description of the topics your clustering discovered along with the 3 most common words from each topic. If the topics do not make sense pick another value of k.

k = 10

In [ ]:
```
centroids, labels, j_hist = kmeans(article_histograms, 10)
```

Out[ ]:
```
(Vector{T} where T[[0.009123992590443903, 0.007348065340944147, 0.004463053
806586829, 0.008436750982367318, 0.004547202552882439, 0.00592714856392926
8, 0.008836275255053415, 0.004393587926192439, 0.001792616132757561, 0.0056
6822029838561  …  0.008436763024849511, 0.008721506614463415, 0.00685931386
9156585, 0.0016119450436204877, 0.02842999315574878, 0.01914417686113878,
0.00035696373459365854, 0.004292075371882438, 0.0014396973048795122, 0.0],
[0.002096434082777143, 0.004186527226047857, 0.009305742458420713, 0.002081
1481049642856, 0.0005496733995221429, 0.0011064227049357143, 0.0015721285581692856, 0.0, 0.0, 0.0029481686680907144  …  0.00048609705917500003, 0.0011952710147035714, 0.01439453692725, 0.00037353848105785715, 0.0, 0.004589352
29269, 0.0, 0.0027596504983871425, 0.0, 0.00084951413388714285], [0.00238215
22092158333, 0.007272858594286666, 0.005623988358019166, 0.0091239212677904
17, 0.00081262476456375, 0.0031566005255558335, 0.004847475540269167, 0.004
4926747941375, 0.003874353425795833, 0.0018031107593533336  …  0.0010673727
292220834, 0.0013835977554208336, 0.003161481654432084, 0.0001094887584387
5, 0.0, 0.0015703772157912499, 0.0010375417017375, 0.0010612656420879165,
0.0, 0.009217060416380833], [0.00860714748152149, 0.004564404633119574, 0.0
04142396061808936, 0.0028325023461619144, 0.002317659421021489, 0.003601956
586131915, 0.004721808267770212, 0.0071336732248314895, 0.031740383354515715, 0.0019510923835676594  …  0.016019775432993405, 0.062387514622761694, 0.
0001759901406319149, 0.019935528310135744, 0.0, 0.01284677553975553, 0.0078
72874322397232, 0.004457060755376383, 0.020495126270557443, 0.0001208342089
0531914], [0.010181037028546154, 0.008600082667953077, 0.004481282442510769, 0.00115947421476, 0.0, 0.0014850330999207693, 0.0, 0.0, 0.001031932386601538
5, 0.004496430910108462  …  0.0, 0.002169056261630769, 0.00041854448232
153847, 0.0, 0.0, 0.006252822506623077, 0.0031408410423038463, 0.0, 0.00156
98581010346153, 0.005148378507007693], [0.005328200052003436, 0.007374816746940937, 0.005777540654170938, 0.0039076321365, 0.0028088888406909377, 0.00
20856555458440623, 0.00193770327245875, 0.0023486990397434374, 0.0, 0.00210
2450189785937  …  0.0004364709617978125, 0.00171766008113, 0.00561198063248
5938, 0.007089661011167812, 0.0, 0.004883919859760937, 0.026984702488657187, 0.0018289967287125001, 0.004463093804159063, 0.0021286019373409374], [0.012617059979489706, 0.007271401618998824, 0.0030273821262382346, 0.00224055
72937588235, 0.006061693793061765, 0.011846573442460885, 0.0046630501848099
99, 0.0032330037988347058, 9.865495664823529e-5, 0.006978972401370589  …
0.015938314066619413, 0.019443829839047053, 0.006380394082399411, 0.0008283
440992167646, 0.011966783366882353, 0.02931718629021853, 0.0001121225825811
7646, 0.01841791966488088, 0.009840558315573527, 0.0008451666328941177],
[0.01240862506231, 0.008106095795872051, 0.008834124164615386, 0.0021088606
52492821, 0.003116278541069487, 0.005520867679143846, 0.001601615157193845
9, 0.002158136550968718, 0.001455958730380769 1, 0.007477085159393 59  …  0.0
024904873934684614, 0.00159549265141, 0.0009450044519699999, 0.000206774379
96846154, 0.0, 0.011902776518204101, 0.0005849128893982051, 0.0022144476923
45641, 0.0003009079630579487, 0.00919699481165154], [0.009858163410996153,
0.01019692844710923, 0.007969244386357181, 0.006317165461352051, 0.00400974
49406282055, 0.0028067573403620515, 0.0007195482647235897, 0.002082839173664103, 0.006064199748164358, 0.0037076436865310257  …  0.003071700700572051,
0.0005816238914205129, 0.005078150969556153, 0.002759937931728205, 0.0, 0.0
```

```
10351790816827178, 0.0005478040193607692, 0.0009342027322041027, 0.00047461
571571794873, 0.002599979933937l8], [0.0014287118363911764, 0.0052190507944
07059, 0.0033570737123517647, 0.002627689374140588, 0.0005294792297658823,
0.0036428632328223526, 0.00038243368185588236, 0.002846960868115294, 0.0026
88684752480588, 0.004464000005134117  …  0.002134641281602941, 0.0035394500
641758825, 0.0038403358471011772, 0.0040888818888500005, 0.0, 0.00136563154
95682353, 0.0, 0.003910197732615294, 0.0027931645643464705, 0.0015527497417
611765]], [4, 8, 8, 4, 3, 3, 3, 9, 10, 9  …  10, 4, 1, 1, 7, 1, 7, 7, 1,
6], [0.8352664006599381, 0.752475725170636, 0.7369837836884727, 0.731211365
6806323, 0.7303354426911302, 0.7298874674214536, 0.7295868000695575, 0.7295
```

In [ ]:
```
article_titles[labels .== 9]
```

Out[ ]:
```
39-element Vector{String}:
 "Anemometer"
 "Anticyclone"
 "Atmospheric pressure"
 "Baroclinity"
 "Barograph"
 "Barometer"
 "Coriolis effect"
 "Dust storm"
 "Erosion"
 "Extratropical cyclone"
 "Flood"
 "Frontogenesis"
 "Hadley cell"
 ⋮
 "Thunderstorm"
 "Tropical cyclone"
 "Typhoon"
 "Weather balloon"
 "Weather forecasting"
 "Weather front"
 "Weather map"
 "Weather modification"
 "Wind chill"
 "Wind direction"
 "Windsock"
 "Wind speed"
```

The clustering has accumulated topics like Anticyclone, Barometer, Weather forecasting, etc. All these titles are related to weather forecasting or prediction. So the cluster seems to be fairly proper.

In [ ]:
```
dictionary[sortperm(centroids[9],rev=true)]
```

Out[ ]:
```
1000-element Vector{String}:
 "weather"
 "wind"
 "pressure"
 "air"
 "tropical"
 "cyclones"
 "winds"
 "cyclone"
 "surface"
 "temperature"
 "storm"
 "fronts"
 "atmospheric"
```

```
⋮
"trainer"
"transmitter"
"treaty"
"unesco"
"unicef"
"unrwa"
"uv"
"van"
"voice"
"voip"
"wireless"
"wto"
```

The 3 most common words found in the article are Weather, Wind and Pressure. All these terms point to association with Weather and climate. Thus we could predict that the model did good.