Pearson Correlation between Annotator 1 and Annotator 2:

t = 5.2223, df = 198, p-value = 4.454e-07

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.2198134 0.4642719

sample estimates:

cor

0.3479432

Pearson Correlation between Annotator 1 and Annotator 3:

t = 1.9997, df = 198, p-value = 0.0469

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.001996246 0.274088921

sample estimates:

cor

0.1406983

Pearson Correlation between Annotator 2 and Annotator 3:

t = 1.6995, df = 198, p-value = 0.0908

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.01915297 0.25441524

sample estimates:

cor

0.1199067

Spearman Correlation between Annotator 1 and Annotator 2:

S = 930970.4, p-value = 1.41e-05

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.3017547

Spearman Correlation between Annotator 1 and Annotator 3:

S = 1162538, p-value = 0.07071

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.1280748

Spearman Correlation between Annotator 2 and Annotator 3:

S = 1268082, p-value = 0.4916

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

0.04891477

Kendall Correlation between Annotator 1 and Annotator 2:

z = 4.1776, p-value = 2.946e-05

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.2347664

Kendall Correlation between Annotator 1 and Annotator 3:

z = 1.7929, p-value = 0.07298

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.1001423

Kendall Correlation between Annotator 2 and Annotator 3:

z = 0.6888, p-value = 0.491

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

0.03835067

Cosine between Annotator 1 and Annotator 2: 0.94

Cosine between Annotator 1 and Annotator 3: 0.91

Cosine between Annotator 2 and Annotator 3: 0.91

Means and Standard Deviations of three annotators:

> mean(ann1)

[1] 4.61

> mean(ann2)

[1] 4.84

> mean(ann3)

[1] 4.51

> sd(ann1)

[1] 1.472666

> sd(ann2)

[1] 1.534986

> sd(ann3)

[1] 1.619572

Cosine between the “mean summary ratings” of four systems:

Cosines between:

System 1 and System 2 = 0.61

System 1 and System 3 = 0.34

System 1 and System 4 = 0.38

System 2 and System 3 = 0.3

System 2 and System 4 = 0.3

System 3 and System 4 = 0.91

This shows that indeed, Systems 1 and 2 are close to each other, and so are Systems 3 and 4. In fact, Systems 3 and 4 are really close to each other.

Cosine between system clusters = 0.38.

Again, this value is much lower than the intra-cluster similarities, which shows that indeed, systems 1 and 2, and 3 and 4, are “closer” within themselves than between.

Summarizability of topics:

Topics in sorted order of summarizability (starting topic index 0):

[6, 41, 7, 12, 42, 27, 49, 28, 10, 11, 34, 15, 18, 21, 43, 16, 8, 33, 22, 14, 17, 30, 38, 29, 13, 24, 46, 31, 40, 4, 39, 37, 9, 19, 1, 5, 25, 35, 45, 47, 36, 20, 44, 48, 23, 32, 2, 26, 3, 0]

By doing paired ANOVA on three annotators - 200 ratings in each (results are in this [spreadsheet](https://docs.google.com/spreadsheets/d/19zE6Dz7bIKZ0TghE_5aQfgebl0xkJ6202Ra_OUGcXDc/edit#gid=379873256)), we see that there is no difference at 5% confidence level (i.e., we cannot reject the null hypothesis). This is because the p-value is slightly higher than 0.05. We did [repeated measures ANOVA](http://blog.excelmasterseries.com/2015/01/single-factor-repeated-measures-anova.html) (“paired” ANOVA) for this, and p-value was computed using this [website](http://graphpad.com/quickcalcs/PValue1.cfm).

*[Since we couldn’t reject the null hypothesis in this case, there is no reason to perform multiple paired t-tests.]*

By doing paired ANOVA on four systems - 50 ratings in each (results are in this [spreadsheet](https://docs.google.com/spreadsheets/d/19zE6Dz7bIKZ0TghE_5aQfgebl0xkJ6202Ra_OUGcXDc/edit#gid=0)), we see that there is in fact some difference (i.e., we can reject the null hypothesis). This is because the p-value is lower than 0.05 (and 0.01). We did [repeated measures ANOVA](http://blog.excelmasterseries.com/2015/01/single-factor-repeated-measures-anova.html) (“paired” ANOVA) as before, and p-value was computed using this [website](http://graphpad.com/quickcalcs/PValue1.cfm).

Then we did multiple two-sample paired t-tests (using Bonferroni Correction with adjusted 𝛼 = 0.05/6 = 0.0083) to see if there is any difference between the four systems. From the [results](https://docs.google.com/spreadsheets/d/19zE6Dz7bIKZ0TghE_5aQfgebl0xkJ6202Ra_OUGcXDc/edit#gid=0), we saw that indeed, Systems 1 and 3, Systems 1 and 4, and Systems 2 and 3 are statistically significantly different from each other (by one-tailed p-value). According to magnitude of the p-value, Systems 1 and 3 are maximally different from each other (lowest p-value), followed by Systems 1 and 4, followed by Systems 2 and 3.

Summarizability of the 50 Hamshahri topics (in sorted order from easiest to summarize to most difficult to summarize):

607-AH 1.0450280435

642-AH 0.724014005464

608-AH 0.709801343789

613-AH 0.606095934029

643-AH 0.59688323115

628-AH 0.545300204577

650-AH 0.545300204577

629-AH 0.531626899517

611-AH 0.490335289054

612-AH 0.484110867769

635-AH 0.47950451633

616-AH 0.449043678592

619-AH 0.429539559602

622-AH 0.398000008634

644-AH 0.354944579066

617-AH 0.322991617507

609-AH 0.307428547319

634-AH 0.299440306929

623-AH 0.221589383467

615-AH 0.208455435022

618-AH 0.191667902344

631-AH 0.172848889229

639-AH 0.112592516392

630-AH 0.105289381877

614-AH 0.0645371280293

625-AH 0.0503244663544

647-AH 0.0168753470219

632-AH -0.0460778088906

641-AH -0.0512235169454

605-AH -0.112412853753

640-AH -0.150861931881

638-AH -0.164806932224

610-AH -0.183625945339

620-AH -0.218421439233

602-AH -0.230476674448

606-AH -0.249567382846

626-AH -0.250646096076

636-AH -0.250646096076

646-AH -0.276913992965

648-AH -0.2912724039

637-AH -0.309138649808

621-AH -0.479864210306

645-AH -0.557030027893

649-AH -0.588175971506

624-AH -0.616333633524

633-AH -0.620128933065

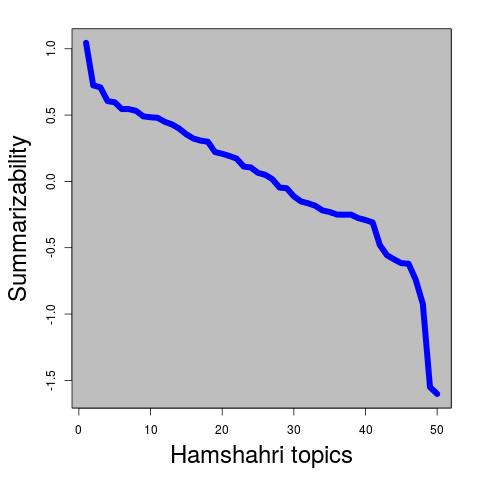
603-AH -0.739543162273

627-AH -0.923256921956

604-AH -1.55078083783

601-AH -1.6023638644

In plot, they look like this:



See that there is a clear trend; a few topics are highly summarizable, a few are highly un-summarizable, and most fall in between.

When I looked into the English description of the topics, here they are (in sorted order of summarizability):

607-AH (Commemorations of Sadi Shirazi)

642-AH (Shahr Theater Programs)

608-AH (House Prices)

613-AH (Children's Rights)

643-AH (Earthquake Damage in Iran)

628-AH (NATO vs. Yugoslavia in 1998)

650-AH (Fluctuations in Gas Imports)

629-AH (Global Drought Predictions)

611-AH (Information Technology and Employment)

612-AH (Internet Users)

635-AH (Iran in 1998 World Cup)

616-AH (Hand-woven Carpet Exports)

619-AH (Iranian Non-oil Exports)

622-AH (Tehran Car Accidents)

644-AH (Electronic Commerce)

617-AH (Tourist Attractions)

609-AH (Fruit Packing)

634-AH (University Acceptance Limits)

623-AH (North Iran Forestry Conservation)

615-AH (Remembrance of Dr Ali Shariati)

618-AH (7 July 1999 Protests)

631-AH (Relations between Iran and the United States)

639-AH (Bovine Spongiform Encephalopathy)

630-AH (Iranian Traditional Celebrations)

614-AH (E-commerce Congress)

625-AH (Places to Visit in Golestan)

647-AH (Buying Military Service Excemption)

632-AH (Olive Oil Benefits)

641-AH (Pollution in the Persian Gulf)

605-AH (Hatamikia's Films)

640-AH (Persian Rugs)

638-AH (Barriers for Investments in Iran)

610-AH (Benefits of Copyright Laws)

620-AH (Freight Transport by Rail)

602-AH (Heart Disease and Smoking)

606-AH (Youth Leisure in Summer)

626-AH (Women in Politics)

636-AH (Air Pollution)

646-AH (Applying to Study out of Iran)

648-AH (Attack on the Twin Towers)

637-AH (Tehran Air Pollution Sources)

621-AH (Television and Mental Health)

645-AH (11 September and Air Travel)

649-AH (Khatami Government Oil Crisis)

624-AH (Films for the Fajr Festival)

633-AH (Daei's World Cup Goals)

603-AH (Gas Rationing in Iran)

627-AH (Nuclear Energy)

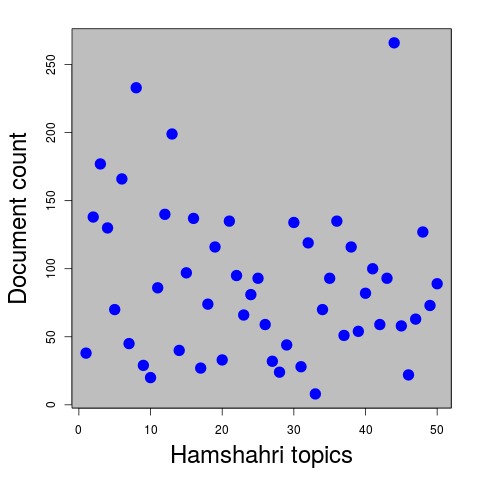
604-AH (Lung Cancer)

601-AH (US Attack on Iran)

Note that the topics, when organized in this order, clearly show a trend: more concrete (and focused) topics tend to be highly summarizable, whereas more abstract (and diverse) topics - also, topics fraught with differing opinions, political controversy, etc - tend to be highly un-summarizable. Also, local and national topics tend to be more summarizable than technical and international topics.

Also note that most topics are local and national, and very few are International. Even fewer relate to the US.

Initially I thought that the less number of documents a topic has, the more summarizable it is. In other words, I thought that there is an inverse correlation between number of documents and summarizability. However, the following plot shows that it clearly is not the case.



This shows a very weak positive (linear) correlation, close to zero:

*Pearson's product-moment correlation*

*data: summ and docc*

*t = 0.6261, df = 48, p-value = 0.5342*

*alternative hypothesis: true correlation is not equal to 0*

*95 percent confidence interval:*

*-0.1931892 0.3593439*

*sample estimates:*

*cor*

*0.08999803*

Hence, the issue of summarizability needs to be investigated further (perhaps by classification based on summarizability?).

Observations from the ROUGE plots:

* Multi-doc results are more pronounced than single-doc results.
* Precision and Recall are following opposite trends (as expected); as compression ratio increases, Recall goes up and Precision goes down.
* With stop words is generally better than without stop words. So we pick with stop words.
* Qualitatively, different ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-3) are similar to each other. So we report ROUGE-3 only.
* Quantitatively, ROUGE-1 is higher than ROUGE-2, which is higher than ROUGE-3 (again, as expected).
* Among binary, tf, tfidf, there is very little difference qualitatively and quantitatively. We pick tfidf.