HW 9 Gabriel Wang

Topic Modelling assists large-scale text analysis. It applies various AI techniques, especially data mining and natural language processing to classify corpus being examined into different topics (Flounas et al). However, this technique was only largely adopted in news area to analyze social contents. For the specific type of corpus being examined in this report, novel, the topic modelling is not expected to work very well. As author of the topic modelling tool we are using mentioned in the documentation, topic modelling has no idea what the topic key words mean (Graham et). Thus, in social content analyzing, topics rarely overlap while overlapping happens a lot in fictions. Thus, this report focuses on identify different story lines of the five books of Song of Ice and Fire (SIF) using topic modelling.

I first separated each book to chapters and did topic modelling of one book at a time. This report only reports the result of the first book – Game of Throne. The remaining four books follows closely the same pattern. The chapter splitting was done via textfilesplitter.com and Game of Throne was separated into 74 different chapters. The first attempt on topic modelling gave awkward result. I pick an arbitral number of topics based on my previous knowledge on the book and the result was reported as follow.

|  |  |
| --- | --- |
| Label | Keywords |
| 0 | JON MAN BACK PAGE EYES BRAN HAND FACE MEN FATHER BLACK LONG TOLD LORD BOY TIME LOOKED ROBB SWORD BROTHER |
| 1 | DANY KHAL DROGO JORAH DOTHRAKI SER VISERYS BLOOD DRAGON KHALEESI ILLYRIO TOLD BROTHER HORSE MIRRI TENT CHILD MAZ HAIR KHALASAR |
| 2 | TYRION LORD CATELYN SER SON LANNISTER LADY TYWIN BRONN LYSA MEN LORDS FATHER RIVERRUN JAIME MAKE EYRIE LANNISTERS RODE HORSE |
| 3 | NED KING ARYA LORD SANSA SER ROBERT JOFFREY QUEEN KING'S LITTLEFINGER PRINCE VARYS SISTER STARK GIRL LADY SEPTA CERSEI GRACE |

Label 1, 2, and 3 make a lot of sense as I can easily identify them as three major story lines. However, label 0 doesn’t make any sense to many, especially the word ‘PAGE’ which I first thought was a character named Page who only makes 1 appearance. Later, I realized that my corpus includes the page number at the bottom of each page…

So, after cleaning up the page number and other confusing element of the corpus, I did the same modelling again and pick the topic number as 4, 7, 10 to examine which topic stands ‘strong’. The result is as follow.

|  |  |  |
| --- | --- | --- |
| LABEL | WEIGHT | KEYWORDS |
| 0 | 0.20194 | DANY KHAL DROGO JORAH DOTHRAKI SER BLOOD VISERYS MEN DRAGON BACK KHALEESI HORSE EYES MIRRI ILLYRIO BROTHER TOLD TENT MAZ |
| 1 | 1.48088 | BRAN EYES ARYA BACK MAN HAND LOOKED FACE SANSA ROBB TOLD MAESTER HEAD BOY TIME FATHER MADE DAY LONG VOICE |
| 2 | 0.76162 | LORD NED SER CATELYN KING ROBERT LADY SON LANNISTER BROTHER KING'S LITTLEFINGER LORDS FATHER MEN HAND STARK RENLY LYSA VARYS |
| 3 | 0.48117 | JON TYRION LORD MAN SER MEN FATHER SNOW BLACK STARK MORMONT BRONN WATCH WALL GHOST BACK HORSE SWORD SAM MAKE |

Table 1: 4-Topics-Mallet

|  |  |  |
| --- | --- | --- |
| LABEL | WEIGHT | KEYWORDS |
| 0 | 0.3108 | JON LORD WALL MORMONT SNOW GHOST WATCH SAM NIGHT'S PYP FATHER BLACK MAESTER COLD BEAR HE'D GRENN JON'S AEMON ALLISER |
| 1 | 0.22011 | TYRION CATELYN BRONN LYSA LANNISTER LADY STARK RODRIK DWARF EYRIE ROCK VARDIS VALE ARRYN ROAD DAGGER MAESTER JAIME MOUNTAIN SKY |
| 2 | 0.34115 | BRAN ARYA ROBB MAESTER FATHER LUWIN WINDOW WOLF SANSA HODOR SUMMER PRINCE RICKON GREY TOWER SEPTA BRAN'S ARMS YARD THEON |
| 3 | 0.25112 | LORD FATHER CATELYN ROBB TYWIN TYRION SON FATHER'S SER LORDS RIVERRUN MEN BATTLE SONS FREY LANNISTERS EDMURE KEVAN BRYNDEN HOST |
| 4 | 0.45646 | NED LORD KING ROBERT SANSA SER QUEEN JOFFREY KING'S VARYS EDDARD LITTLEFINGER CERSEI GRACE PYCELLE RENLY STARK THRONE COUNCIL BARRISTAN |
| 5 | 0.11779 | DANY KHAL DROGO JORAH DOTHRAKI VISERYS BLOOD DRAGON KHALEESI MIRRI ILLYRIO MAZ SER TENT DUUR KHALASAR DROGO'S DAENERYS IRRI STALLION |
| 6 | 2.27681 | BACK MAN SER EYES HAND MEN TOLD FACE LOOKED TIME LONG BOY HEAD MADE SWORD THOUGHT BROTHER BLACK VOICE ASKED |

Table 2: 7-Topic-Mallet

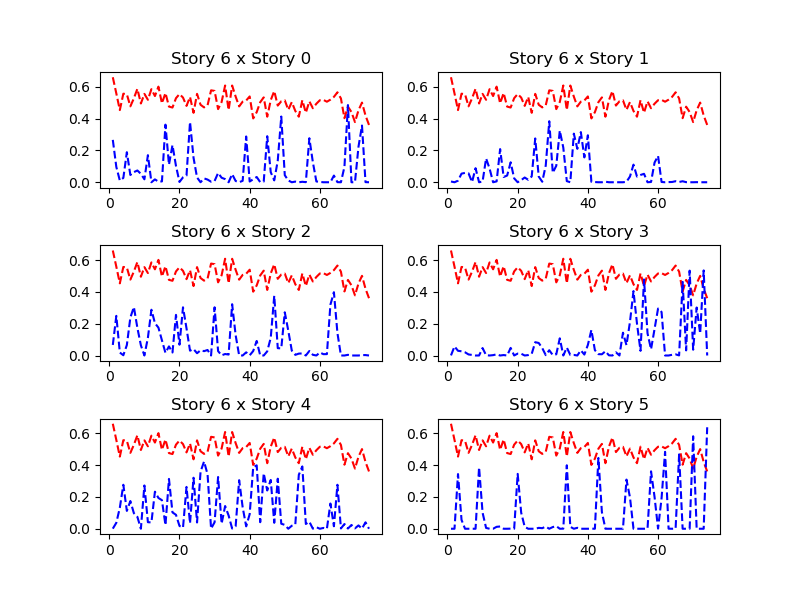
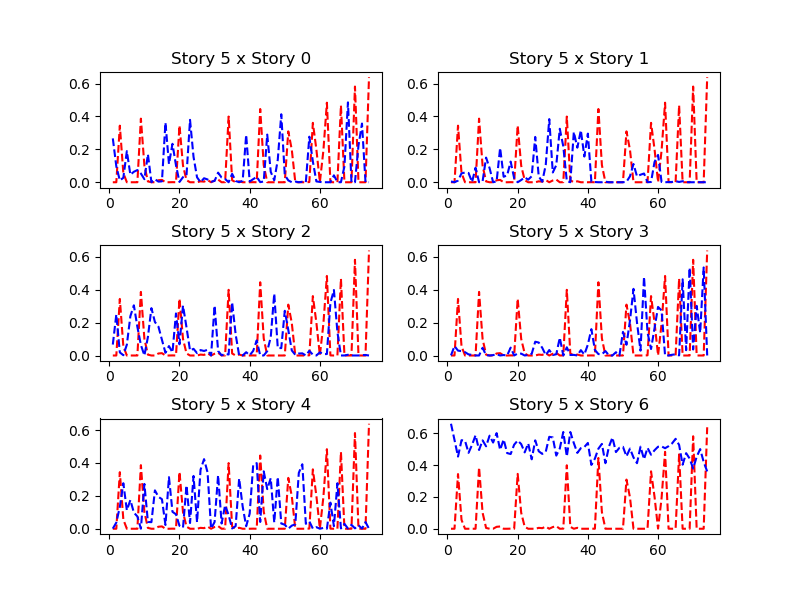
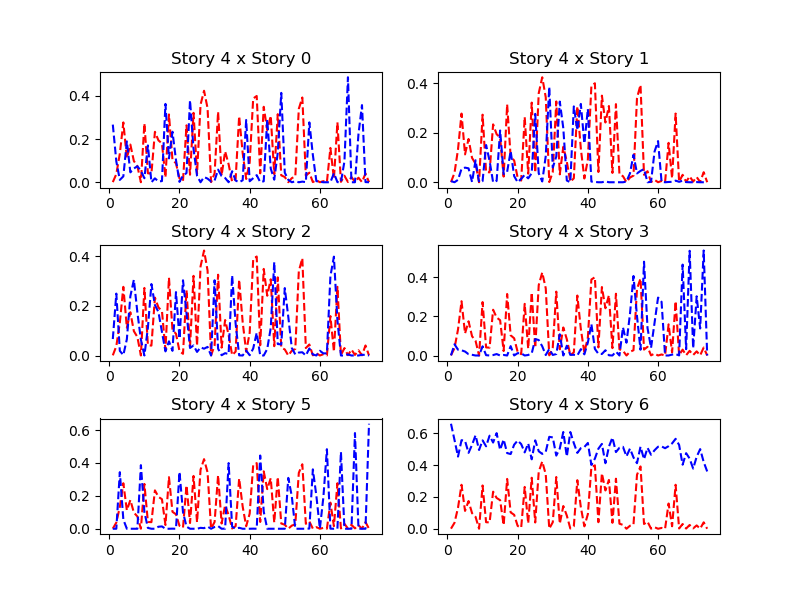
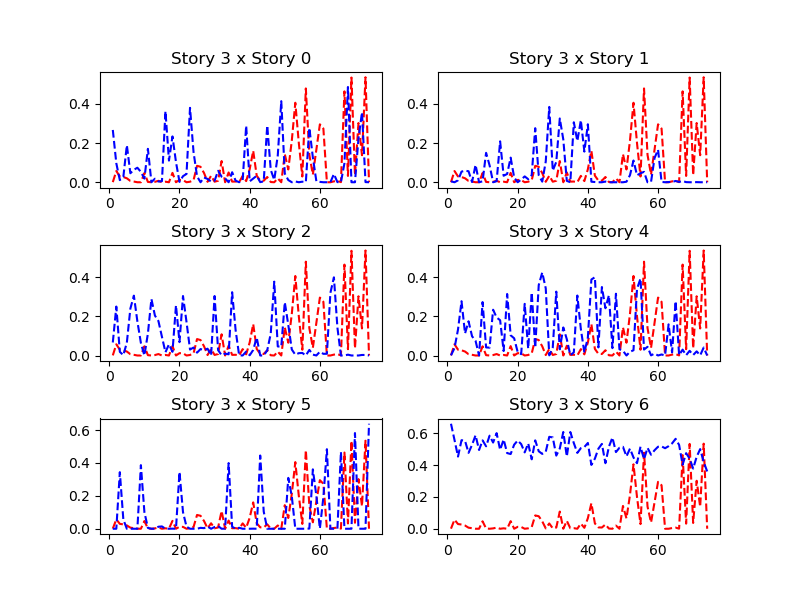
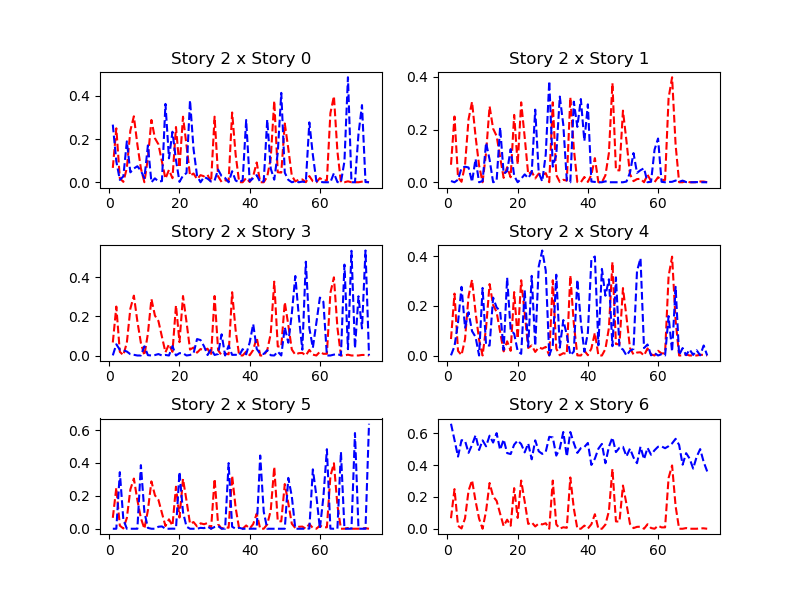
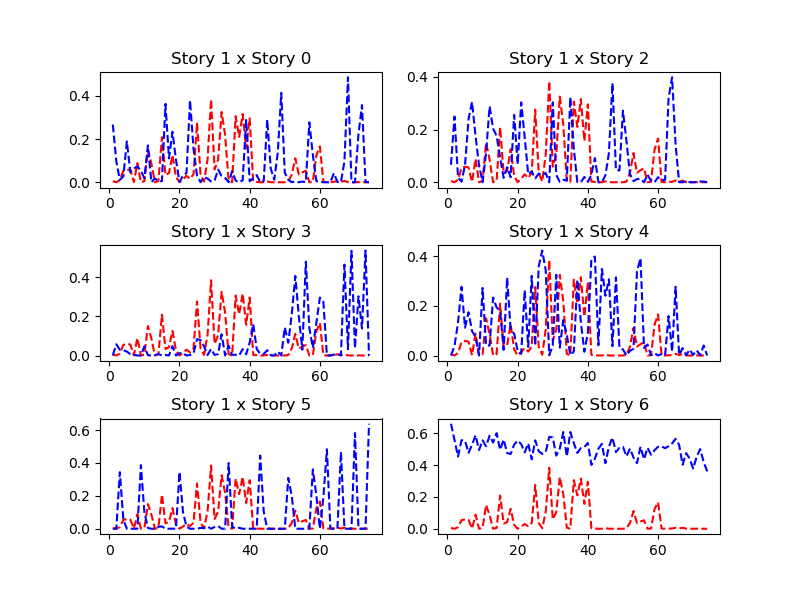
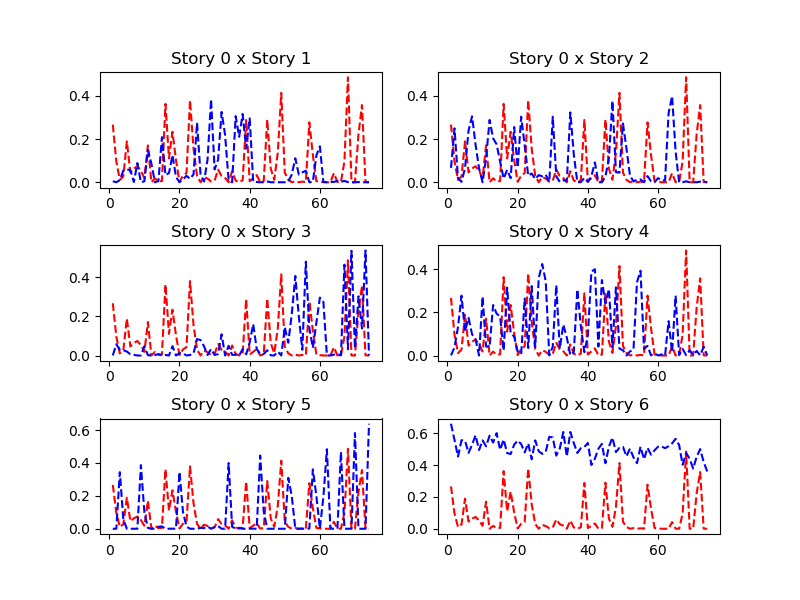
|  |  |  |
| --- | --- | --- |
| LABEL | WEIGHT | KEYWORDS |
| 0 | 0.24155 | ARYA SANSA JOFFREY SEPTA QUEEN FATHER PRINCE LADY MORDANE SYRIO KING WINDOW CLOAKS SHE'D WOLF GOLD JEYNE HOUND FAT NEEDLE |
| 1 | 2.5751 | BACK MAN EYES MEN HAND TOLD FACE BOY TIME BLACK LONG LOOKED HEAD MADE THOUGHT LORD VOICE SWORD LEFT DEAD |
| 2 | 0.34723 | SER KNIGHT STEEL ARMOR SHIELD HELM SWORD CLOAK ROYCE SILVER HORSE ARM METAL COLD GARED FELL ARROWS BLADE SWORDS WAYMAR |
| 3 | 0.10606 | DANY KHAL DROGO JORAH DOTHRAKI VISERYS SER DRAGON BLOOD KHALEESI MIRRI ILLYRIO BROTHER MAZ DUUR KHALASAR TENT DAENERYS STALLION DROGO'S |
| 4 | 0.20916 | CATELYN LADY RODRIK SER LYSA EYRIE SISTER YEARS ARRYN VALE VARDIS SINGER UNCLE ROOM LITTLEFINGER RIVERRUN LANNISTERS TULLY HUSBAND ROAD |
| 5 | 0.28897 | BRAN ROBB MAESTER LUWIN SUMMER FATHER HODOR RICKON GREY JON CHILDREN WOLF THEON BRAN'S GREYJOY DIREWOLF WINTERFELL NAN FATHER'S OSHA |
| 6 | 0.46459 | NED LORD KING ROBERT SER KING'S VARYS STARK CERSEI LITTLEFINGER EDDARD GRACE QUEEN MAESTER PYCELLE THRONE RENLY JAIME COUNCIL REALM |
| 7 | 0.20878 | LORD FATHER SER ROBB CATELYN SON TYWIN FATHER'S LORDS RIVERRUN BROTHER FREY EDMURE SONS MEN LANNISTERS BATTLE HOST KEVAN WALDER |
| 8 | 0.17738 | TYRION STARK BRONN LANNISTER DWARF GOLD JAIME TYWIN ROCK SHAGGA MORD FIRE TYRION'S ROAD AXE CASTERLY WORD CROWS CLANSMEN MOON |
| 9 | 0.22374 | JON LORD WALL SNOW MORMONT GHOST SAM WATCH MAESTER NIGHT'S PYP GRENN BEAR ALLISER AEMON BROTHERS BENJEN JON'S HE'D BASTARD |

Table 3: 10-Topic-Mallet

The two strongest topics are the topic with key words “DANY KHAL DROGO JORAH …” and “JON LORD WALL SNOW MORMONT”. They are the stories of Dany Targaryen and Jon Snow, the leading characters of the whole book, arguably the representation of Fire and Ice in the title. These story lines are largely separated from every other line. Jon’s story took place at the Wall, the northernmost part of Westeros while Dany’s story took place in Essos, a totally different continent. Dany’s story line only overlapped other story lines for once in the whole book one. And Jon’s story crosses the story line ‘the Life of Winterfell’ at the beginning of the book for about 5 chapters. Another pattern noteworthy is the topic 1 in 4-topic modelling and topic 6 in 7-topic modelling with keywords, they have large portion of overlapped keywords and the early one actually ‘absorbs’ topic 6 and topic 2 into it. Furthermore, this kind of absorption is also found when comparing 7-topic-modelling and 10-topic-modelling. All topics except the two strong-standing and the relatively-strong-standing story line (topic 2) in 4-topic modelling can be absorbed into topic 3. And no matter how many topics we picked, the keyword ‘VOICE, LONG, SWORD’ is still there, this may worth a further look in the future. However, this isn’t the main feature of topic modelling we want to explore today. We summarizes the story lines as follow based on content of the corpus.

|  |  |
| --- | --- |
| Label | Summary |
| 0 | the Wall (main character: Jon, Mormont, Sam, Aemon, Pyp, the Watch …) |
| 1 | the Trial of Tyrion (main character: Tyrion, Catelyn, Lysa, Bronn, the Vale …) |
| 2 | the Winterfell (main character: Bran, Arya, Robb, Luwin, Sansa, Hodor, …) |
| 3 | the Riverland (main character: Catelyn, Robb, Tywin, Edmure, Frey ...) |
| 4 | the Kingslanding (main character: Ned, the King, Sansa, Cersei, Joffery …) |
| 5 | the Dragon Queen (main character: Dany, Khal Drogo, Jorah, Viserys, Dragon …) |
| 6 | other (no main characters) |

 Let’s further take a look at when the two strong story lines cross other story lines to get some quantitative proof why these two story lines always stand strong. We used the 7-topic-modellng to plot our story line. And for now, we chose to ignore topic 6 which absorbs other potential topics for a better visualization. A python script was used to clean up the data, the script was attached.

The plots of each story line cross other story lines are as follow: 

The x-axis is the chapter index and the y-axis is the percentage of the story line takes in the chapter. The way to tell if two story line crosses is to find the peaks of the two story line appears together. I attached a more inclusive figure of all seven story lines and the result is as follow.

Remember that the two ‘strong’ story lines we want to inspect are story 0, ‘the Wall’ and story 5, ‘the Dragon Queen’. Take a closer look at the plots of story 0 and story 5. The plot confirms that the story ‘the Wall’ crosses the story ‘the Winterfell’ at the beginning of the book while ‘the Dragon Queen’ never crosses any of the other stories.

This could lead to some very interesting and useful result. The novel will undoubtedly converge, that is all story lines will finally become one and that is when the peak of the story arrive. However, how Martin is going to finish the story lines is unknown. A more detailed comparison of the topics of the five already exists book will give us a better answer.