발표 대본 가안

1. 인사 다함께 : 안녕하세요 팀 알파911 입니다. 팀 이름은 알파코 9기 1번째 팀프로젝트 1조 라는 의미로 지었습니다. 원활한 발표를 위해 QnA는 발표 마지막 시간에 진행하도록 하겠습니다.

2. 발표 순서 설명

3. 프로젝트 개요

저희는 프로젝트 주제 선정을 위해

분석할 데이터가 충분히 있는 사이트를 팀원 각자 탐색하고,

각 사이트를 어떻게 분석하여 어떤 인사이트를 뽑아낼지 각자 제안서를 작성했습니다.

그리고 이 과정에서 크롤링 복습심화가 가능한지도 추가적으로 검토했습니다.

다양한 창의적인 프로젝트 제안 중

다수의 팀원 의견을 반영하여 봄툰 구독자 분석을 하기로 결정 했습니다.

분석을 통해 구독자 관심과 심리를 파악하여 새로운 인기 웹툰을 만들기와 굿즈 제작을 비롯해

비인기 작가들에게 개선 방향 인사이트를 제공할 수 있을거라 생각했습니다.

4. 프로젝트 예상 시나리오

가장 먼저 어떤 결과물을 도출할지 어디서부터 시작해야할지

<예상 시나리오>를 의논 했습니다.

큰 절차 항목은 필요할만한

1. 봄툰 데이터 크롤링

2. 수집한 데이터 전처리

3. 데이터 시각화

4. 데이터 분석

이렇게 4단계로 정의 했습니다. 각 단계별 세부항목은 화면을 참고해주시기 바랍니다.

5. 프로젝트 실제 시나리오

실제 프로젝트 진행하며 큰 절차 항목은 동일했으나

예상 시나리오와 달라진 부분들이 있었습니다.

크롤링 방법에선 <셀레니움>을 활용하기도 했고

전처리 과정에서는 영문화 처리를 하기도 했고

시각화 과정은 matplot 그래프 생성 대신 LDA를 생성했습니다.

6. 프로젝트 수행 WBS

절차에 대한 WPT(Work Breakdown Structure) 그래프입니다.

저희 팀은 협업 도구로서 실제 현업에서 많이 쓰이는 GitHub를 적극 활용했습니다.

7. 프로젝트 수행 WPT

마찬가지로 WPT 입니다. 화면 참고 부탁드립니다.

8. vmfhwprxm xla

우리 알파911 팀은 모든 기능 구성을 팀원 모두가 만들어봤습니다.

각자 개성을 살려 기능 구현 코드를 작성하고

팀원 서로에게 자신의 코드를 가감없이 보여주었습니다.

이렇게 다양한 방식의 코드를 서로가 경험하고 성장하도록 진행 했습니다.

취합한 코드중 프로젝트에 적합한 코드는 무엇일지 서로 의논하며 Main 코드를 선정했습니다.

그렇기에 우리팀은 모두가 적극적인 리더쉽을 가진 팀장으로서 수평구조 팀이라는 떳떳한 장점을 갖게 됐습니다.

9. 봄툰 웹 분석

그럼 실제 시나리오 순서대로 과정을 발표하겠습니다.

우선 화면과 같이 웹 페이지 분석을 했습니다.

봄툰은 모든 페이지가 JSON으로 되어 있었기에 크롤링 난이도가 있습니다.

10. 크롤링 코드

직접 각자 개개인이 크롤링 해본 결과 401권한에러 404페이지에러 405요청방식에러 크롤링차단에러 코랩에러 등 다양한 에러를 경험했습니다.

또한 크롤링 시도하려는 페이지 마다 다른 JSON에서 참조하는 값들이 있었고

봄툰 서버의 보안설정으로 인해 웹 분석과 코드 복잡도가 상당했습니다.

그러나 과정을 서로 공유하며 에러를 해결했고

이 과정에서 수업에서 배우지 않았던 셀레니움도 스스로 학습하기도 했습니다.

결과적으로 모두가 각자의 방식으로 크롤링에 성공했고 .

의논을 통해 메인 코드를 선정해 툰 데이터를 다양하게 추출했습니다.

11. 크롤링 결과 데이터

고생 끝에 수집한 데이터 결과 화면 입니다.

12. 결측치 제거 + 댓글 코퍼스 생성

수집한 데이터에서 불필요한 결측치를 제거하고

빠른 분석을 위해 댓글만 남기도록 처리한 코드 입니다.

이렇게 얻어낸 기본 댓글 데이터셋을 기본 코퍼스로 사용하기로 했습니다.

13. 댓글 코퍼스

앞서 보여드렸던 처리한 결측치 제거 후 얻은 코퍼스 결과입니다.

14. 불용어 사전 구축

코퍼스를 보고 불용어를 수작업으로 추가하기도 하고

온라인에서 제공하는 한국어 불용어를 탐색해 적절히 추가하기도 했습니다.

15. 한국어 분석기 라이브러리 성능 비교

다음으론 한국어를 어떻게 분석할 것인지 의견을 나누었고,

각각의 성능의 장단점(속도, 정확도 등)을 꼼꼼히 비교하여 최적의 라이브러리를 활용하려고

머리를 맞대며 고민했습니다.

konlpy를 비롯한 다양한 한국어 분석기를 사용해보기로 했습니다.

16. 불용어처리

우선 분석기 사용 전 불용어를 처리를 먼저 처리 후 코퍼스를 확인해봤습니다.

이 때 상당한 양의 코퍼스와 불용어 리스트를 비교 처리하면 시간소모가 컸습니다.

여러가지 분석기의 성능을 비교 하기에 많은 시간이 걸리겠다는 생각에

어떻게 속도 개선을 할 수 있을지 일상에서 꾸준히 고민하고 수업을 돌이켜 봤습니다.

그 결과 numpy의 array 특성을 떠올려 직접 코드에 적용해 봤고

시간이 약 만 배 가까이 빠르게 처리 됨을 확인하였습니다.

17. 이후 okt 명사 추출 + 불용어 처리도 시도하고

18. hannanum 명사 추출 +불용어 처리도 시도하고

19. mecab 명사 추출 + 불용어 처리도 시도하고

20. khaiii 명사 추출 불용어 처리도 시도해보고

21. pororo 명사 추출 불용어 처리도 시도해봤습니다.

22. WordCloud 한글 깨짐 처리 + 워드 클라우드 생성

처리된 코퍼스가 한글이라 워드클라우드 한글깨짐 설정 후 워드 클라우드를 생성을 했습니다.

23. 각 WordCloud 결과

각 워드클라우드 결과 입니다.

불용어 처리만 한 워드클라우드 부터

순서대로 메캅, okt, 한나눔 분석기 까지 적용한 워드클라우드입니다.

한나눔은 왜인지 현재 머릿속을 들여다 보는듯한 워드클라우드가 나왔네요.

뽀로로와 카이 분석기는 버전 호환성 에러가 있고

해결하기에는 다른 일도 있기에 다음에 따로 분석해보기로 하며 보류했습니다.

그 외 기타 분석기들도 마찬가지로 보류를 하게 되었습니다.

워드 클라우드만 봐서는 단어의 크기가 크지만

실질적으로 얼만큼의 빈도가 있는지 확인할 수 없다는 점을 캐치했습니다.

24. pyLDA 코드

그래서 LDA를 이용해 워드 빈도수 확인과 각 워드들의 상관관계율을 분석해보았습니다.

25. LDA 시연 영상

코드로 생성한 LDA 영상 입니다.

26. 코퍼스 한글 -> 영어 전환

앞서 분석에 대한 결과물을 통해 팀원분들과 함께 조금 더 다르게 인사이트를 얻을 수 있진 않은지 고민해보았고,

시나리오 구상 때 언급 되었던 플랜B를 적용해 보기로 했습니다.

플랜 B는 코퍼스를 영문으로 변경 후

토큰화와 불용어 처리를 하면 쉽고 효율적이지 않을까에서 비롯되었습니다.

처음엔 엑셀의 한영전환 기능을 이용하면 빠르고 간편하지 않을까 싶었습니다.

우리의 코딩 창작 능력 발달을 위해 가급적 파이썬 코드로 처리하기로 했습니다.

하지만 대량의 코퍼스 약 7만대 이상의 문장을

무료로 번역 가능한 라이브러리를 찾는데 부터 난관이 시작됐습니다.

찾아 보던 중 OpenNMT라는 라이브러리는 한국어 데이터와 영어데이터 두 가지를 가지고 지도학습 시키는 모델이어서 바로 사용할 수 없었습니다.

여러 탐색 후 googletrans의 Translator를 찾게 되어 사용했습니다.

하지만 구글 트랜스 라이브러리도 큰 코퍼스를 처리하기엔 메모리 사용률이 매우 높아

중간에 에러가 발생하였고

해결 방법을 모색하여 결국 코퍼스를 나누어 번역 처리하는 것으로 해결했습니다.

27. 한국어 코퍼스 영문 전환 결과

번역 결과 입니다.

화면과 같이 한국어 문맥의 100% 일치하지 않게 번역되었습니다.

해당 라이브러리는 한국어 단어의 의미가 분명할수록 번역의 정확도가 올라갔습니다.

1번째 줄에서 '아니'감탄사를 no로 번역한 것을 제외하고는 의미가 크게 손상되지 않았습니다.

허나, 3번째 줄의 '쓰다'라는 뜻은 영어의 write으로 쓰인 게 아니라 use로 쓰였음을 알 수 있습니다.

28. 영문 1만 워드클라우드

앞서 처리한 영문 코퍼스 1만개에 사이킷런 불용어를 적용해 생성한 워드 클라우드입니다.

워드클라우드를 분석해보니 im 과 ha tt와 같은 단어들은 사이킷런에서 제공하지 않음을 확인했습니다.

29. 영문 전체 워드 클라우드 + 불용어 추가

이어서 처리가 필요한 불용어를 사이킷런 불용어와 함께 추가하여 7만여 전체 영문 코퍼스로 워드클라우드를 생성한 화면입니다.

보다 나은 결과로 보여집니다.

30. 영문 버전 LDA

한국어 LDA와 마찬가지로 영문 코퍼스에 대한 LDA도 생성했습니다.

31. 기술 스택

지금까지 팀 프로젝트를 진행하며 사용한 각종 기술스택입니다.

32. 자체평가

이것으로 알파911 팀의 발표를 마치겠습니다.

들어주셔서 감사합니다.