①변경된 점들

1. 블로그의 텍스트 파싱 기능 추가

2. 범용성을 높힌 정규식으로 교체

3. URL 읽어오기 기능을 추가하기 위한 밑작업

②추가사항

1. 텍스트 파싱 + 2. 범용성 높은 정규식

범용성을 높힌 정규식 덕분에 필요없는 html 속성들도 같이 딸려 파싱된다.(속도 저하에 원인)

pics += re.findall('<img src="(.\*?)" (.\*?) class="se-image-resource" />', html)

sticker += re.findall('<img src="(.\*?)" (.\*?) class="se-sticker-image" />',html)

text += re.findall('<span(.\*?)>(.\*?)</span>',html)

유형 1(이걸 기준으로 한다 하면)

<span style="" class="se-fs- se-ff- " id="SE-8bc8444a-b906-4fe9-97a0-39e64fba4e67">열심히 사진을 찍었습니다^-^ㅎㅎ</span>

유형 2(class와 id가 다르고)

<span style="" class="se-fs- se-ff-nanumbarungothic " id="SE-28b8065b-db89-4e33-99d5-89521d79ff2b">그림말풍선은 2017년 초에 아이디어가 떠올랐고</span>

유형 3(아마도 다른곳에서 ctrl+c v 해보이는 텍스트들은 세부 설정 없이 그냥 span태그만 있음)

<span>파이썬 공식홈페이지에 설치형 버전이 있던데, 그림에 표시한거 다운받아서 설치했다.<br></span>

이런식으로 블로그마다 조금씩 차이가 나서 모든 유형을 포함하는 정규식을 사용하다보니 쓸대없는 se-ffnanumb 같은 코드들도 같이 딸려옴