Created by Firejune at 2016/10/26 , Last modified 2016/10/27

60여개 언어를 지원하는 OCR 엔진 Tesseract.js

Tesseract.js는 C++로 작성된 Tesseract OCR 라이브러리를 자바스크립트로 포팅한 것으로 텍스트의 방향을 자동으로 탐지하며, 단락을 구분해 내거나 단어 및 문자의 경계를 탐지하는 등의 인터페이스를 제공합니다. Emscripten을 이용하여 tesseract.js-core라는 이름으로 포트 되었으며, 이 자바스크립트 파일의 용량이 무려 2.7MB에 달합니다. 브라우저의 리소스를 많이 잡아먹어서인지 WebWorker에서 동작하게 되어있고 Node에서는 `child_process` API를 이용하는군요. 코어와 언어별 트레인드(Trained) 데이터는 최초 인식 때 한 번만 가져오고 이후부터는 캐시에서 불러옵니다.

한글의 인식률이 어떤지 궁금해서 원래 있던 데모에 한글모드를 추가해 보았습니다. 고딕 계열 폰트로 작성된 한글 이미지의 인식률이 가장 높았으며, 불분명하다고 판단하는 경우가 아주 많았습니다. 상단 탭을 한글로 맞추고 한글이 들어간 이미지 파일을 드롭하면 다른 파일의 인식 테스트도 가능합니다.

```
Tesseract.recognize(myImage)
         .progress(function  (p) { console.log('progress', p)    })
         .then(function (result) { console.log('result', result) })
```