맞춤법 검사 엔진을 사용하여 다중 언어의 OCR 인식 정확도를 높이기

개요

아래는 영어, 러시아어, 중국어(간체), 일본어 및 한국어로 게티즈버그 연설의 일부를 발췌한 PDF 입니다. 이 문서는 개발자/사용자에서 Advantage OCR 엔진의 성능과 그 한계를 보여줍니다.

English

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

Russian

4 счета и 7 лет тому назад наши отцы принесенные вперед на этот материк, новую нацию, понятые в вольности, и предназначенные к предложению что все люди будут созданным равным. Теперь мы включены в большом гражданской войне, испытывая ли та нация, или любая нация поэтому после того как я поняты и так после того как я предназначены, могут длиной вытерпеть. Мы встречаны на большом battle-field того войны. Мы пришли предназначить часть того поля, как окончательное отдыхая место для тех которые здесь дали их жизни которые та нация могла жить. Altogether fitting и правильно что мы должны сделать это.

Chinese Simplified

四个比分和七年前我们的父亲带来在这个大陆，一个新的国家，设想在自由和致力提议人生而平等。 现在我们参与一次巨大内战，测试那个国家，或者任何国家，因此设想和如此致力，是否可能长期忍受。 我们在那场战争一个伟大的战场遇见。 我们来致力那个领域的部分，作为一个最后的休息处为这里给他们的生活那个国家也许居住的那些人。 它是一共贴合和适当的我们应该做此。

Japanese

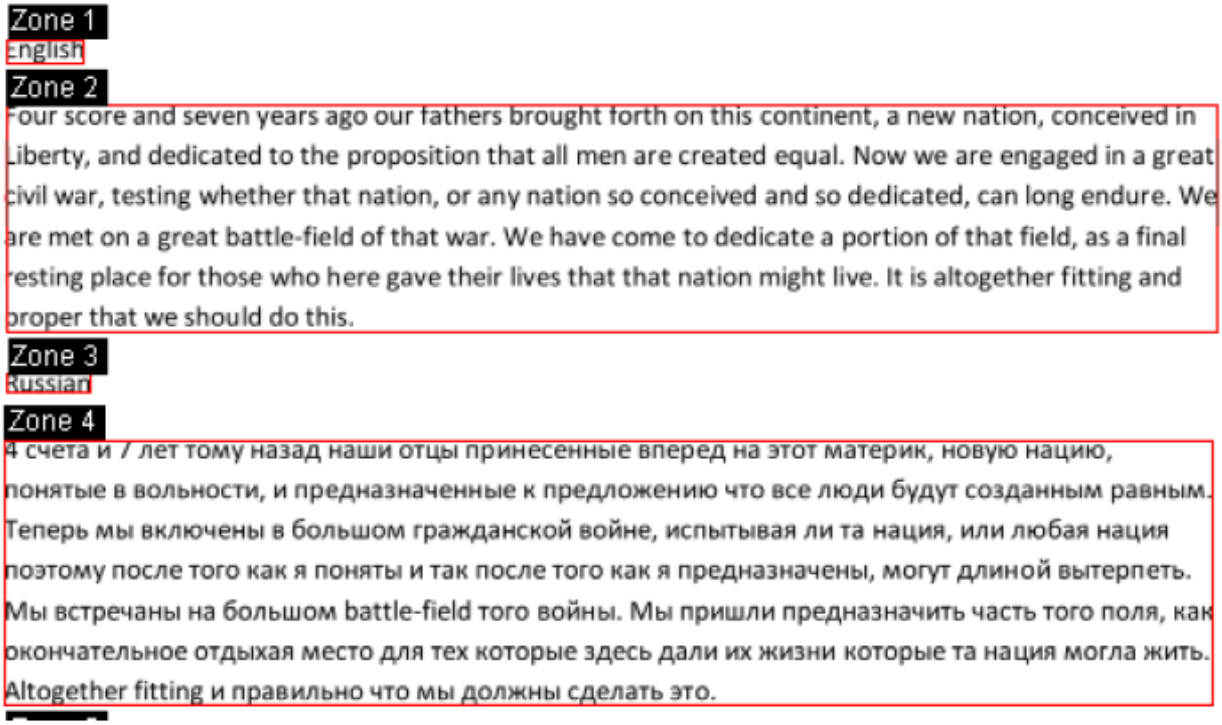
すべての人が作成された同輩であること 4 つのスコアおよび 7 年前にこの大陸、新しい国家で持ち出され、自由で想像され、そして提案に捧げられる私達の父。 今度は私達はその国家、か国家が従って想像され、そう捧げられて、長く耐えることができるかどうかテストする大きい内戦で従事している。 私達はその戦争の大きい戦場で会う。 私達はここにその住む国家がかもしれない彼らの生命を与えた人のための永眠の地としてその分野の部分を、捧げることを来た。 私達がこれをするべきであることは全体で適切、適切である。

Korean

4 개의 점수 및 7 년 전에 이 대륙, 새로운 국가에 생기고, 자유에서 생각되고, 건의안에 모든 사람은 평등하게 창조되었다 바쳐지는 우리의 아버지. 지금 우리는 저 국가, 또는 아무 국가나 그래서 생각하고 이렇게 바쳐, 오랫동안 영속할 다는 것을 시험하는 중대한 남북 전쟁에서 접전된다. 우리는 저 전쟁의 중대한 전장에 만나진다. 우리는 여기에서 살지도 저 국가가 모른 그들의 생활을 준 그들을 위한 마지막 휴게실로 저 분야의 부분을, 바치기 위하여 왔다. 우리가 이것을 해야 한다 전부 적당하고 적당하다.

OCR은 일반적으로 인식 정확도가 낮은 순위의 문자들에 대해 "빈칸을 채워"로 언어 사전들을 참조합니다. 예를 들어, "Will"이란 단어에 대해, OCR은 마지막 문자가 "1" 인지 또는 "I" 인지를 결정하기 위해 사전에서 "Wil1" 또는 "Will" 중 어떤 단어가 있는지를 확인.참조합니다.

Advantage 엔진이 자동으로 하나의 이미지(텍스트, 테이블 및 그래픽을 찾는)를 영역화할 때, 통상 하나의 영역내에 텍스트의 블록들을 그룹핑합니다. 사용자는 문서내에서 각각의 언어가 별도의 단락이 되며 별도의 텍스트 블록임을 볼 수 있습니다. 다시말해, 텍스트의 각 블록은 하나의 별도 영역이 됩니다.



OCR이 텍스트의 언어를 식별하려고 시도할 때, 영역마다 하나의 언어를 사용합니다. 그래서 PDF의 경우, OCR은 각 영역에 포함된 언어를 식별하려고 시도합니다.

Advantage 엔진은 일반적으로 맞춤법 검사 엔진으로 알려진, 언어 사전들로 세가지의 다른 시스템들을 지원합니다.
특히, Hunspell은 가능한 많은 다른 언어 사전들을 갖고 있는 오픈 소스 맞춤법 검사 엔진입니다.

Advantage 엔진내에 Hunspell을 지원하기 위해 필요한 라이브러리들을 획득하고 사용자가 지원을 계획하는 언어들에 대해 필요한 언어 사전들을 획득함으로서, 사용자는 Advantage 엔진이 영역마다 각각의 언어를 식별하려고 시도하게끔 할 수 있으며, 이로서 인식 결과의 정확도를 높일 수 있습니다.

Hunspell 맞춤법 검사 엔진 및 언어 사전을 획득하기

Hunspell 맞춤법 검사 엔진을 획득하고 이를 Advantage 엔진을 사용하는 사용자의 개발 환경에 설치하는 방법은 아래의 온라인 문서에서 찾을 수 있습니다:

https://www.leadtools.com/help/leadtools/v19/dh/fo/leadtools.forms.ocr~leadtools.forms.ocr.ocrspellcheckengine.html

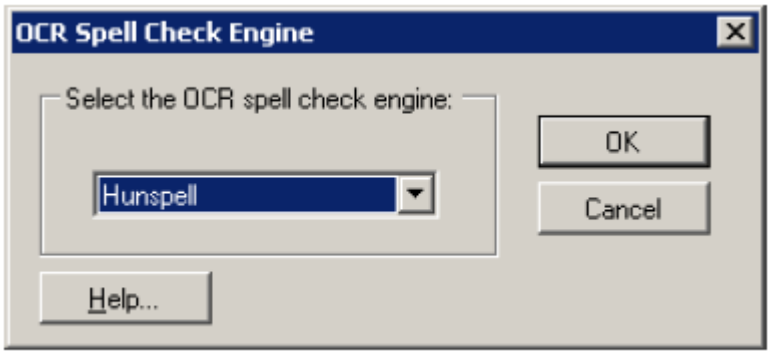
OCR 다중 엔진 데모내에 Hunspell 맞춤법 검사 엔진을 사용하기

이미지를 로드

1) "₩LEADTOOLS ??₩Bin₩Dotnet4₩Win32₩CSOcrMultiEngineDemo_Original.exe" 데모 프로그램을 실행
2) 시작에서, LEADTOOLS Advantage 엔진의 사용을 선택
3) 파일 | 종료 메뉴 항목에서 기본으로 로드된 이미지를 삭제
4) 파일 | 열기 메뉴 항목에서 PDF 셈플을 로드

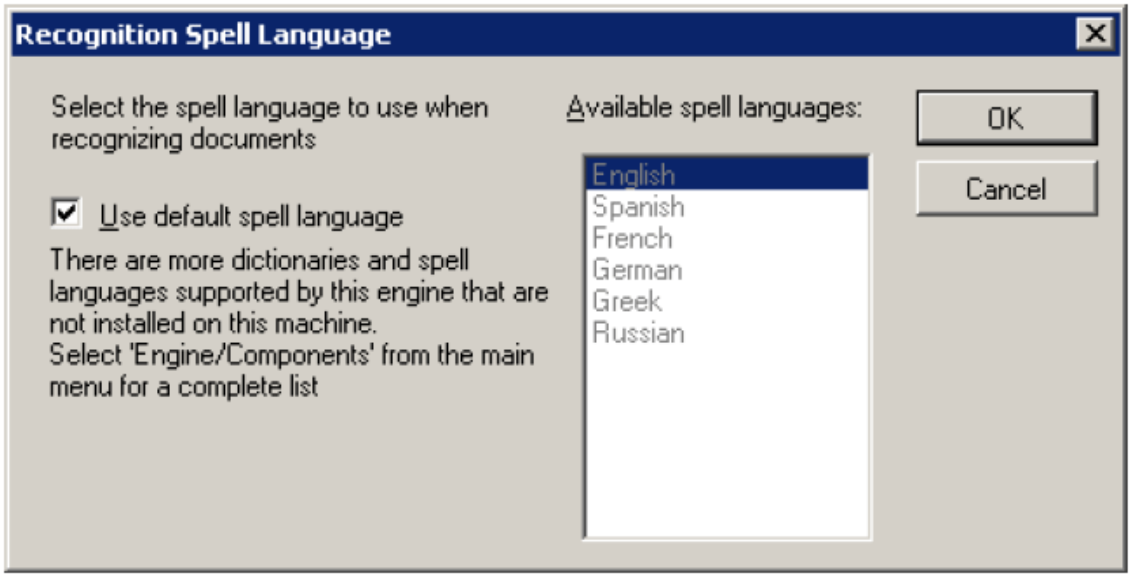
Hanspell 맞춤법 검사 엔진을 선택하기

5) OCR | 맞춤법 검사 엔진 사용..: 하나의 선택 대화 상자를 표시하는 메뉴 항목
6) 드롭 다운에서, "Hunspell" 맞춤법 검사 엔진을 선택



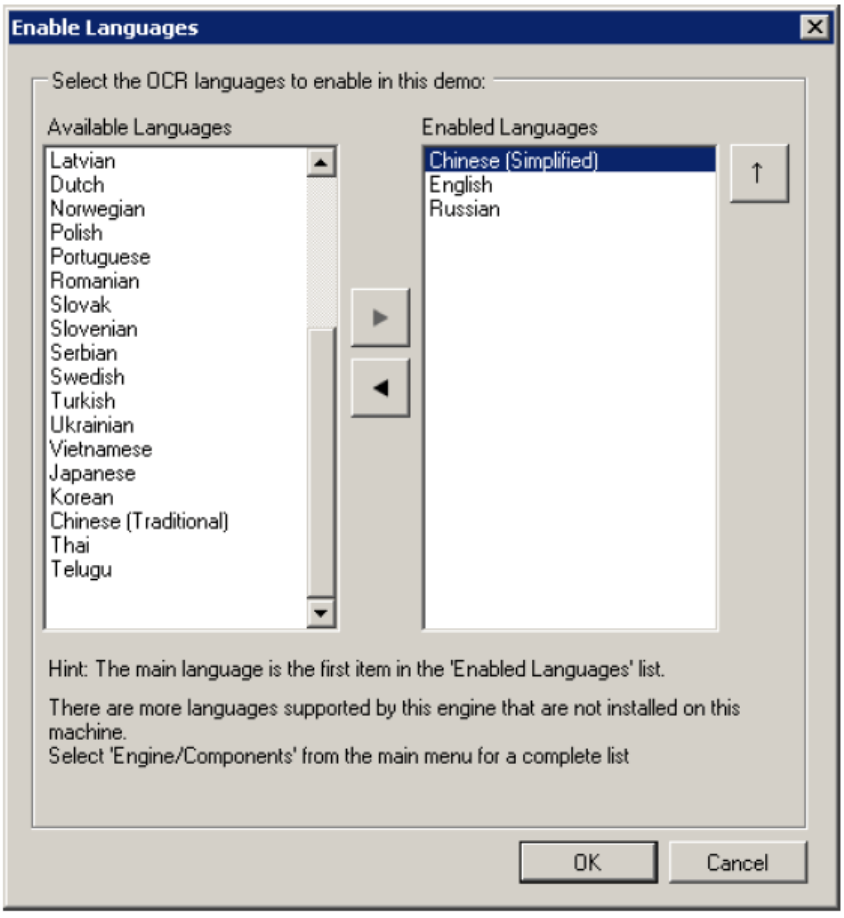
언어 사전 목록

7) OCR | Spell Language 사용..: Hunspell 맞춤법 검사 엔진 사용이 가능한 모든 언어 사전들의 목록을 포함하는 대화 상자를 표시하는 메뉴 항목



각 영역에서의 검색을 위해 Advantage 엔진의 사용 가능한 언어들

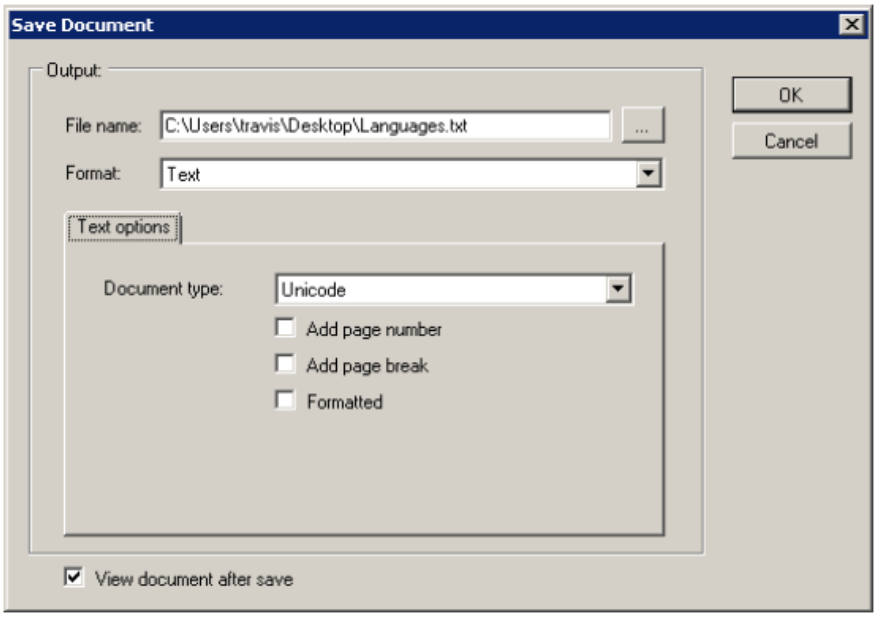
8) 엔진 | 언어들 사용...: 각 영역에서의 검색을 위해 Advantage 엔진에 대하 사용자가 원하는 언어들을 선택하는 대화상자. 셈플 문서의 경우, 언어들이 순서대로 있음을 확인합니다.



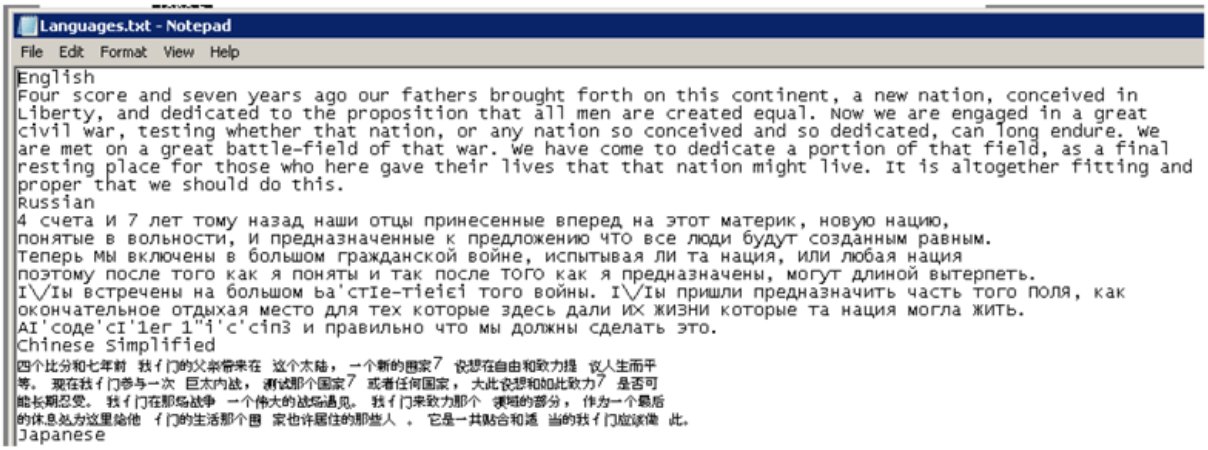
문서를 인식하고 그 결과를 저장

9) OCR | 문서 저장 : 문서를 인식하고 그 결과를 저장하는 메뉴 항목

10) 유니 코드 인코딩 텍스트로 결과를 저장

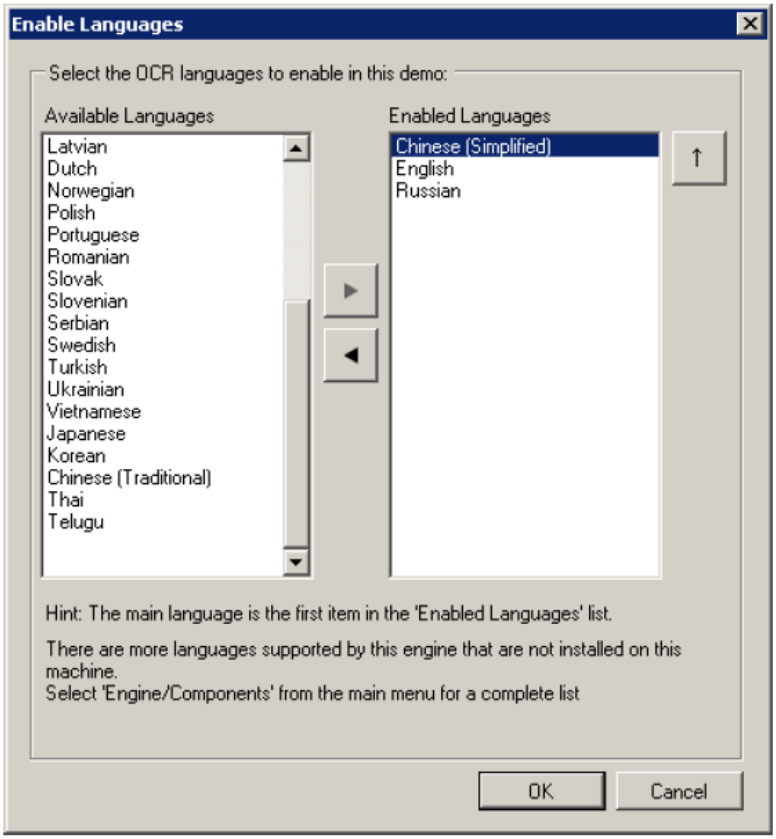


결과



제한

한 영역의 언어를 판별할 수 없는 경우, Advantage 엔진은 디폴트 언어를 선택할 수 밖에 없습니다. 디폴트 언어는 사용 가능한 언어들 대화상자에서 선택 가능한 언어들 목록의 첫번째 언어가 됩니다:



대부분의 아시아 언어들은 언어 사전을 갖고 있지 않습니다. 따라서 Advantage 엔진이 하나의 아시아 언어를 식별할 수 있도록 하는 것은 가능하지 않습니다.

이 이미지의 경우에, 사용자는 문서에서 발견된 세가지 아시아 언어들의 하나만을 성공적으로 인식할 수 있을 것입니다.

그러나, 다른 두개의 언어들(영어와 러시아)은 가용한 언어 사전들을 갖고 있기 때문에, 이들은 높은 정확도로 식별 및 인식되었습니다.